Identification of subjects from reconstructed images

Identification of individual subjects based on image reconstructions generated from fMRI brain scans

Arthur Mercier[1]

Supervisor: Xucong Zhang[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Arthur Mercier
Final project course: CSE3000 Research Project
Thesis committee: Xucong Zhang, Nergis Tömen

Abstract

Reconstructing seen images from functional magnetic resonance imaging (fMRI) brain scans has been a growing topic of interest in the field of neuroscience, fostered by innovation in machine learning and AI. This paper investigates the possible presence of personal features allowing the identification of subjects from their reconstructed images. Identifying the extent to which personal information is present is necessary to prevent privacy and data protection breaches. Additionally, personal features may reveal information about how people see the world, furthering work in computer-brain interfacing or helping people with neurological conditions that affect sight. In this paper, a CNN model is presented that allows to identify subjects from their reconstructed image with an average accuracy of 90.4%. An encoder-decoder model [1] was used to produce the reconstructed images from the Generic Object Data set. The accuracy shows that personal features are indeed present in the reconstructed images, raising important ethical and legal considerations when using image reconstruction technology.

Keywords: Visual stimulus reconstruction, fMRI, Generic Object Decoding, Subject identification, Personal data, Privacy

1 Introduction

The human brain and how it encodes visual data has been an area of research for a long time. With the recent developments in machine learning, researchers have found ways to train computational models to read and decode visual data using brain imaging techniques. One of the main examples of this research is recreating seen images from the brain activity detected by functional magnetic resonance imaging (fMRI) scans. This entails using machine learning to reconstruct the natural image the person was seeing at the time the scan was taken from the fMRI scan recordings of their brain activity. Researchers have already presented models designed to achieve this objective [2]. However, the reconstructed images are still perfectible and further research is needed to accurately categorise them. The question that will be considered in this paper is **the identification of individual subjects based on reconstructed images**. We will more particularly explore to what degree it is possible to identify the subject whose brain scan was used as input to produce a given recreated image.

Privacy and the protection of medical data is a growing concern as we continue to move towards a more data-driven world. It is thus important to consider the ethical issues linked to the technological advances in machine learning. The brain activity recorded by fMRI scans consists of personal data. Information collected from the reconstructed images should therefore be carefully considered. Future users and suppliers of this technology will have to take appropriate measures to ensure data privacy.

Furthermore, mapping the reconstructed image to a user could reveal personal ”signatures” particular to each participant with respect to their reconstructed images. Those signatures would correspond to features that appear in every image generated from a specific user's brain scan. This could further our understanding of how a given individual sees an image, advancing work in computer-brain interfacing or helping people with neurological conditions that affect sight.

This paper is structured as follows. Chapter 2 will go over the methodology of the research, explaining the process, models and data-sets used to identify the subject from a reconstructed image. Chapter 3 will present the results of this research, which will be analysed and discussed in chapter 4, comparing the result differences based on model, subject and input image. Chapter 5 will expand on the ethical aspects of this topic and the importance of responsible research. Finally, chapter 6 will conclude with a summary of the findings of the analysis and will point to further research that could be explored in future works.

2 Deriving subjects from reconstructed images

2.1 Background information

To be able to start identifying subjects from reconstructed images, it is first necessary to compute those reconstructed images. Several models have been developed to accomplish this task, each with different implementations, strengths and weaknesses. However, they all follow the same general process as seen in Figure 1. Firstly, the subject is shown a set of different images, representing a range of things that vary in format, shape or colour. In this paper we will focus on natural images, namely real pictures of objects or animals found in nature, thus generating scans that resemble day to day brain activity. While the subject is viewing the input image, their brain activity is recorded with the use of functional Magnetic Resonance Imaging. fMRI machines measure the small fluctuations in blood flow that occur with brain activity. By measuring these fluctuations in the visual cortex, the part of the brain that receives and processes the visual information received from the eyes, we can record the brain activity prompted by the input image. This data is then given to one of the decoder network models mentioned above, which will draw out latent features from the brain activity to then derive the original image input the subject was seeing when the brain scan was taken.

2.2 Identifying subjects

We will propose a model allowing to determine if it is possible to identify a subject, using an image reconstructed from that subject's brain activity. The process that will be used to achieve this is shown in Figure 2. The reconstructed image will be passed to a Convolutional Neural Network (CNN) which will then classify it as belonging to one of the

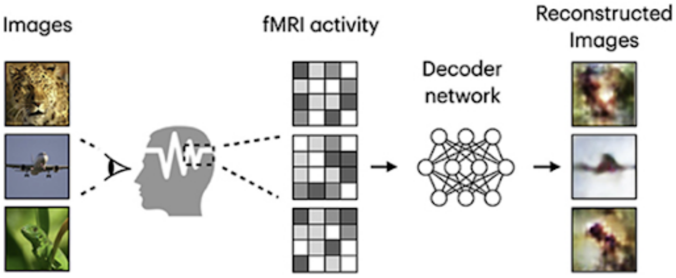


Figure 1: Functional Magnetic Resonance Imaging (fMRI) scan of the subject's brain activity is taken while they look at an input image. A network model then decodes the features of the scan and reconstructs the input image. Image sourced from Rakhimberdina et al [2].

possible subjects. A CNN was chosen as model due to its ability to detect both low and high level features.

The strength of the CNN comes from its convolutional layers which filter the pixels in the inputted images with their surrounding neighbouring pixels, ending up with an abstracted version of the features we started with. Stacking multiple of these layers on top of each other allows the CNN to capture low level features (details) with the lower layers and higher level features (ex: shape, image composition, position) with the higher layers. This results in the neural network being able to detect both patterns that are grouped together and patterns that are spread around the whole image. Dropout layers are placed in between the convolutional layers to prevent overfitting. Overfitting emerges when the model is trained too specifically on the features present in the training data, resulting in accurate predictions for the training data but not any new data. The dropout layer randomly turns off a certain frequency of nodes by setting their input to 0. This stops all neurons from synchronously optimizing their weights and converging on the same features, thus decorrelating their weights and stopping the model from overfitting on specific features.

An added benefit of a CNN model architecture is that it resembles the biological hierarchy of the visual cortex, the same place the brain activity is taken from, thus reproducing the hierarchical low and high levels in the brain into a high and low level structure in the neural network [3]. This could also help track the individual signatures that can be used to identify a subject.

2.3 Reconstructed images

Different models based on different neural network architectures have been constructed for image reconstruction. The main distinction between them is the difference between non-generative and generative models. Non-generative models use only the input from the fMRI, essentially creating a mapping from the brain activity to the different features that it will construct in the image. Non-generative models, as used by [4] [1] [5], can be viewed as complex functions, outputting results based on the input. These are shown to yield accurate object shape and position in the reconstructed images [2].

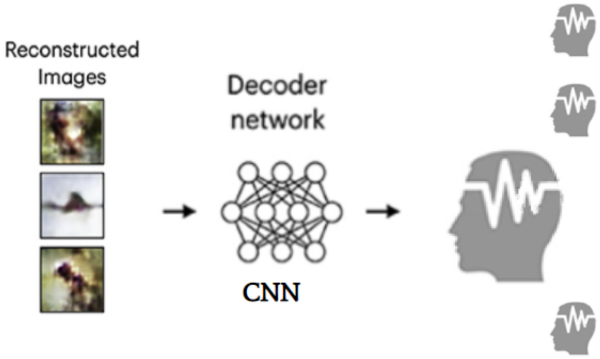


Figure 2: A CNN decodes the personal signatures found in the reconstructed images and classifies the input as one of the possible subjects.

Generative models, as used in [6] [5] [7] [8] [9], assume that data is generated from a probability distribution. They therefore work to identify this probability distribution that the sample inputs are based on. Generative models are comprised of a generator and discriminator.The generator generates new sample images based on this distribution whereas the discriminator is trained to differentiate between real samples and generated samples. If the discriminator can not tell the difference between the generated and real sample, it is outputted.

Images are reconstructed by generative models by first training the model on natural images. The input fMRI images containing the brain activity are then used as constraints, restricting the feature space from which the model can generate only to images that could produce this brain activity. The advantage of this method is that generated images are made to look like natural images, thus making them clearer and more recognisable to a human [2]. However, as they are not generated solely from the input data, the reconstructed image may contain additional semantic features that were not present in the originally viewed image. For this reason, generative models were not selected for this paper. The use of non-generative models minimises the chance of fake features or generated content to obfuscate possible personal features that could be used to identify subjects.

The model used in this paper is a non-generative, encoder-decoder model [10]. This model is made up of an encoder network and a decoder network as shown in Figure 3. The encoder is trained to map stimulus images to corresponding fMRI activity. Conversely, the decoder goes through a supervised training to map fMRI activity to corresponding images. Combining both back to back creates a self-supervised network allowing for the decoder to be trained on unlabelled data. Both the encoder and decoder also use convolutional and pooling layers, gaining the benefits of a CNN. This model, which is the most recent non-generative model available, was chosen because it improves or outperforms

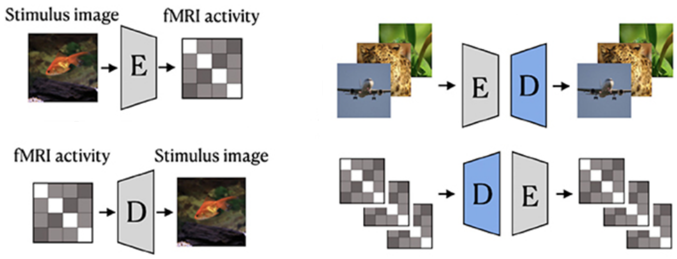


Figure 3: Supervised training of encoder and decoder, left. Self-supervised training of decoder using fixed trained encoder, right. Image sourced from Rakhimberdina et al [2].

other available models. It was presented as an improvement to Beliy and al's[4] model mainly through its perceptual similarity loss function [11], outputting results that were better at capturing high level features such as object shape, background and colour. This model was also shown by [2] to outperform the non-generative model proposed by [5].

The data used to train the model was sourced from the Generic Object Decoding (GOD) data-set [12] and ImageNet's Large Scale Visual Recognition Challenge (ILSVRC) [13]. These datasets are widely used in the natural-image reconstruction field. They were thus chosen to minimize confounding variables when comparing research papers. The model had also been built and tuned for this data-set, thus ensuring better results. Both data sets and model are available for non-commercial use.

2.4 Measuring retained personal features

Once the reconstructed images have been obtained, they are imputed into the CNN to be classified according to their respective subjects. The accuracy of the CNN classification will help determine whether personal features are present in the image. If the CNN can accurately detect which subject was used to reconstruct a given image, features must exist in the reconstructed image to serve as basis for the CNN to differentiate. It is to be noted that the CNN's accuracy not being significantly higher than random selection, does not necessarily mean that no personal features are present in the reconstructed images. It just signifies that they could not be captured by the CNN model.

If the accuracy of the CNN model was not significantly higher than the accuracy of random selection of subjects then, by extension, comparing accuracy over different subjects and image inputs would not be relevant. However, if the accuracy of the CNN model was significantly higher than the accuracy of random selection of subjects, implying the presence of personal features in the reconstructed images, comparing the accuracy across the different subjects and images would be warranted. Some people might be easier to recognise due to some particular attributes or just because they possess more diverse personal features than the rest of the subjects in the model's perspective. Additionally, some seen images might make subject classification easier, because the personal features are more easily captured by the CNN in their reconstructed version.

3 Results

3.1 Reconstructed images

The model used to produce the reconstructed images was Self-Supervised Depth Reconstruction model proposed by Gaziv et al [10]. The source code, setup and parameters used are available at their repository 'SelfSuperReconst'[1]. The image and fMRI data sourced from the Generic Object Decoding (GOD) data-set [12] and ImageNet's Large Scale Visual Recognition Challenge (ILSVRC) [13] are also available on the repository. This data consists of 1200 training images and 250 test images over 5 subjects and 50 prompt images. Once trained, running the model on the test dataset produced 250 reconstructed images containing one image per subject per unique prompt image.

Figure 4 presents a sample of the reconstructed images from subject 3 together with their prompt image. The reconstructed images produced during training and testing are shown on the left and right respectively. Figure 5 presents the quality accuracy and average rank score of the reconstructed images for each subject during a 5-way and 10-way comparison. During a n-way comparison, the reconstructed image is compared to n candidate images including the original prompt image. Accuracy denotes the similarity accuracy of the reconstructed image to the prompt image using Perceptual Similarity metric [11], a CNN that calculates image precision in a similar way to human perception. To derive rank score, candidate images are ranked by similarity to the reconstructed image using the Perceptual Similarity metric. The average rank score is the average rank of the prompt image. For example, an average rank of 1.5 could signify the prompt was ranked first for half of the comparisons and second for the remaining half. The average accuracy between all subjects was 74.4% for the 5-way comparison and 66% for the 10-way comparison. The average rank between all subjects was 1.45 for the 5-way comparison and 2.05 for the 10-way comparison.

3.2 Subject Identification

The subject identification model was developed using Keras[2], a library that provides abstraction for the layers in a neural network. It was made of 4 layers. Firstly, the reconstructed images were imputed in a 2-dimensional convolutional network layer, composed of 32 filters with 3x3 kernel size. Secondly, a dropout layer was placed. After passing the result through a *relu* activation function, they entered a second identical convolution layer. Finally, they were passed to a dense interconnected layer using a softMax activation function that resulted in the end class probabilities. This architecture tested the best, additional layers did not produce any improvements. Furthermore, the model uses the Adam optimization algorithm. The learning rate and steps per epoch hyper-parameters were fine-tuned using grid search

[1] https://github.com/WeizmannVision/SelfSuperReconst
[2] https://keras.io/

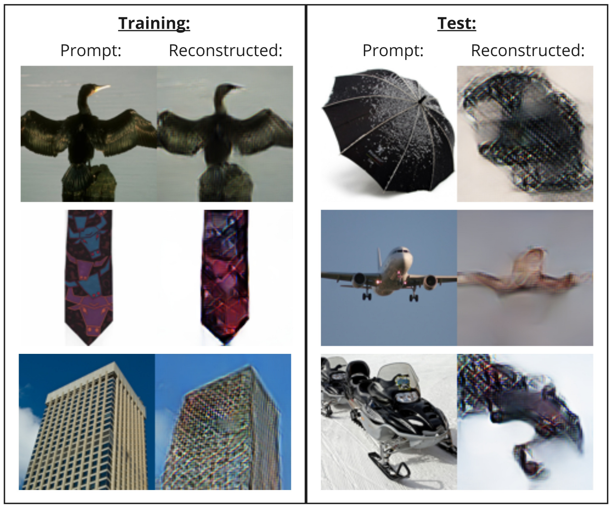


Figure 4: Prompt image next to reconstructed image after passing through model. Training samples are shown on the left and test samples are shown on the right

algorithm.

The reconstructed images were partitioned into an 80% training data and 20% test data. This separation was done according to prompt images, resulting in 40 reconstructed images for each subject based on the same 40 prompts being used as training data (200 images total) and 10 reconstructed images for each subject based on the same 10 prompts being used as testing data (50 images total). The data was divided this way to eliminate possible bias that would result from different image features used to train recognition on different subjects.

As a first experiment, the model was trained with the prompt image and the reconstructed image (see figure 4). The second experiment used only the reconstructed images. The first experiment would allow the model to also capture correlating features between the prompt image and the reconstructed image. These correlating features pertain to how the reconstructed image changes based on the prompt image and subject. A hypothetical example would be, if subject 3 sees a plane, they tend to view the windows as smaller. The second experiment abstracts out those correlating features. Hypothetically, plane shaped reconstructed images from subject 3 tend to have smaller windows. Although only using the reconstructed image gives the model less information, any resulting accuracy would only be derived from features present in the reconstructed images. The context of this second experiment would be more relevant to privacy attacks than the context of the first experiment. The second experiments simulates the conditions of an attacker that has no access to the prompt image, only to the brain scan or reconstructed image.

Experiment 1: Reconstructed images with prompt image

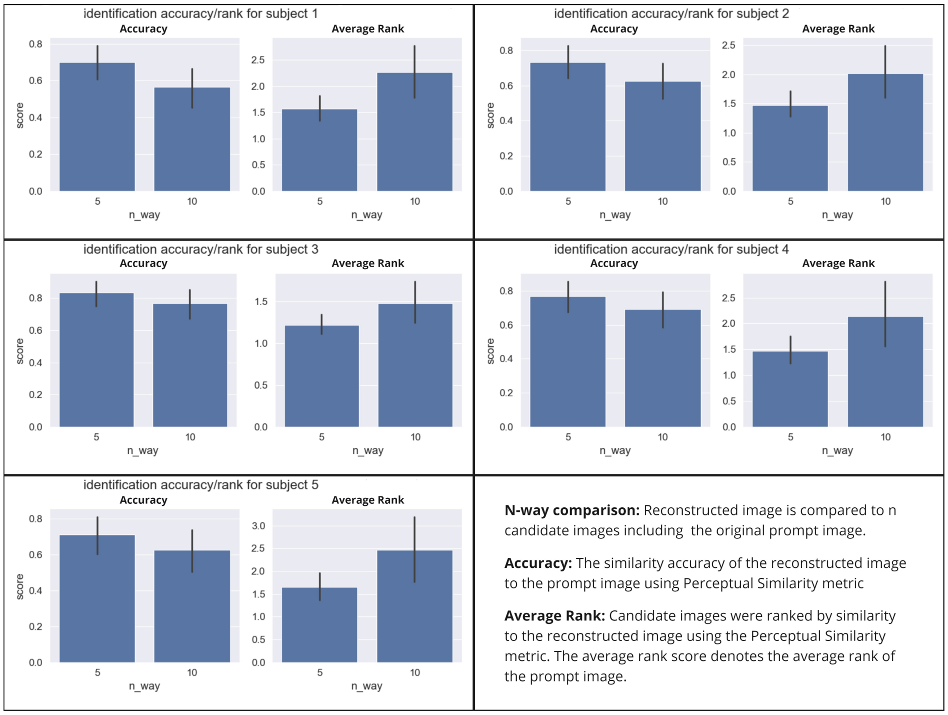
Grid search resulted in a learning rate of 0.0001, 85 epochs, 3 steps per epoch and a drop frequency of 1%. For this experiment, an extra convolutional layer and dropout layer were added before the dense layer as it resulted in a better performance. This was due to the prompt image being added, thus doubling the input size. The overall identification accuracy over the 5 subjects, averaged over 10 runs, resulted in 90.6% with a highest accuracy run of 96%. Figure 6 shows the individual accuracy per subject. The accuracy is shown using three metrics: precision, recall and f1-score. Precision illustrates the ability of the classifier to not label a negative sample as belonging, calculated with: True positive / (True positive + True negative). Recall conversely refers to the model's ability to classify all positive samples correctly, calculated with: True positive / (True positive + False Positive). F1-score is a harmonic mean of precision and recall. Figure 7 shows the accuracy changes per different image prompt.

Experiment 2: Reconstructed images only

Grid search resulted in a learning rate of 0.00015, 85 epochs, 3 steps per epoch and a drop frequency of 1%. The overall identification accuracy over the 5 subjects, averaged over 10 runs, resulted in 90.4% with a highest accuracy run of 96%. Figure 8 shows the individual accuracy per subject. The accuracy is shown using the same three metrics; precision, recall and f1-score. Figure 9 shows the accuracy changes per different image prompt.

4 Discussion

The results found in chapter 3 show that a CNN model is able to identify individual subjects based on their reconstructed images with an average accuracy of 90.4%. This signifies that personal information remains in the reconstructed images. This information could have been introduced in three different ways during the image reconstruction process. The first consists of the different ways subjects see images. The image in the subject's visual cortex being the one reconstructed, any personal features resulting from how the subject views the image will remain in the reconstruction. The second way is from the fMRI scan itself. The scan measures the oxygen content and flow in the brain which is then used to simulate brain activity. This means the scan contains a multitude of information such as the brain's architecture or the organisation of brain activity, which is different in every brain. This information could have remained in the reconstructed images and have been used to re-identify the subjects. The last way is from the neural network used for reconstruction. Neural networks have different strategies to process different input, this might have a noticeable effect on the result. The model might reconstruct images differently based on the subject's brain, resulting in different reconstructed images. Identifying the strategy used could enable to identify the subject. All three ways contain private information and thus pose a privacy risk to the users of this technology.



N-way comparison: Reconstructed image is compared to n candidate images including the original prompt image.

Accuracy: The similarity accuracy of the reconstructed image to the prompt image using Perceptual Similarity metric

Average Rank: Candidate images were ranked by similarity to the reconstructed image using the Perceptual Similarity metric. The average rank score denotes the average rank of the prompt image.

Figure 5: Quality accuracy (higher is better) and average rank score (lower is better) of the reconstructed images for each subject during a 5-way and 10-way comparison.

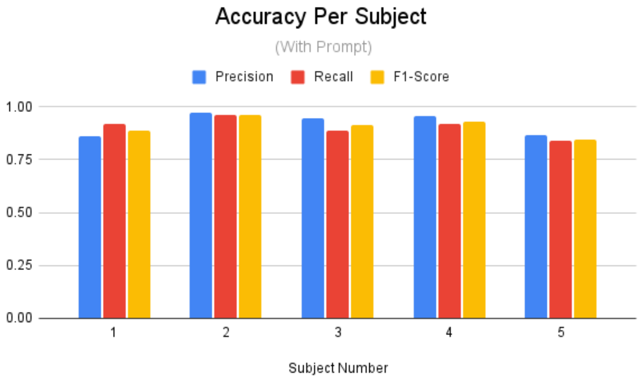


Figure 6: The accuracy per subject is shown using three metrics; precision, recall and f1-score. Precision illustrates the ability of the classifier to not label a negative sample as belonging. Recall conversely refers to the model's ability to classify all positive samples correctly. F1-score is a harmonic mean of precision and recall.

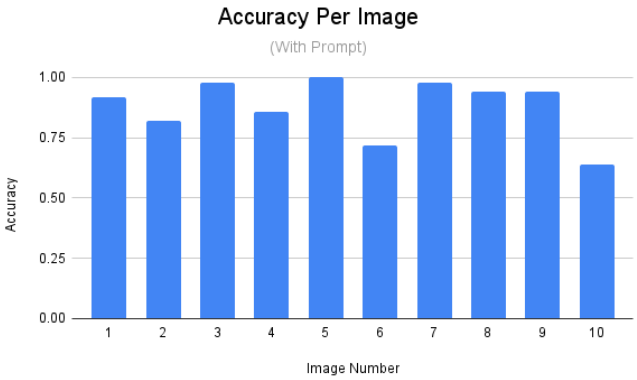


Figure 7: The figure denotes the accuracy with which the CNN was able to classify reconstructed images to the correct subject for each of the 10 input images.

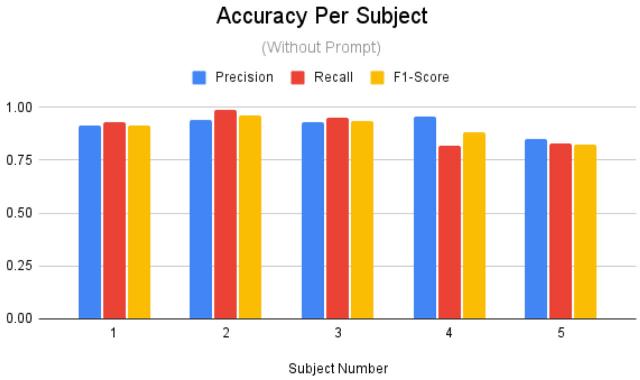


Figure 8: The accuracy per subject is shown using three metrics; precision, recall and f1-score. Precision illustrates the ability of the classifier to not label a negative sample as belonging. Recall conversely refers to the model's ability to classify all positive samples correctly. F1-score is a harmonic mean of precision and recall.

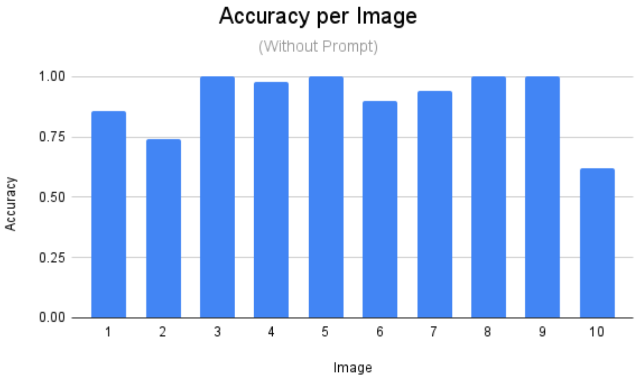


Figure 9: The figure denotes the accuracy with which the CNN was able to classify reconstructed images to the correct subject for each of the 10 input images.

4.1 Quality accuracy of the reconstructed images

The quality of the image reconstructions was measured with accuracy and rank, using the Perceptual Similarity metric [11], scoring the same way as human perception. The image reconstructions have an average quality accuracy of 74% and a rank of 1.45 for the 5-way comparison and 66% and a rank of 2.05 for the 10-way comparison. However, not all of the images have the same accuracy. Images reconstructed from subject 3 are shown to be the best for all metrics. Conversely, subject 1 has the lowest image accuracy while subject 5 has the worst image rank. This gives a quality order of 3 followed by 2 and 4, followed by 5 and 1.

A general trend can be observed when comparing the image reconstructions quality per subject with the subject identification accuracy per subject. Subjects who have higher quality image reconstructions tend to have a higher classification accuracy returned by the CNN model. The images from subjects 2,3 and 4 who are shown to be of better quality than images from subjects 1 and 5 also have a better identification accuracy. This signifies that higher quality images are better at containing personal features that could help re-identify subjects. Identification is not fully linked to the quality of the image however. Even though 3 has a better quality, 2 performs best most consistently. Additionally images from subject 1 perform better than 4 in some cases (91.6% to 88.2% for f1-score in experiment 1).

4.2 Analysing accuracy difference between experiments

The first experiment with prompts resulted in an average accuracy of 90.6 % whereas the second experiment with prompts resulted in an average accuracy of 90.4%. The difference between the two is of only 0.2%, which is not significant enough to be marked as a definite improvement. As stated in the results section 3, an improvement in accuracy when including prompts would show the presence of transitional personal features between the prompt and the reconstruction. This is because the added prompt image is the same for all subjects and thus does not contain any features the CNN can use to differentiate. As there has been no accuracy increase, we know that the CNN was not able to capture any transitional personal features that could help to further identify subjects. This does however not mean that they do not exist. It just means that either the CNN model was not able to capture them, or that they did not provide identification information other than the personal features already included in the reconstructed image. From a data protection perspective, this means that attackers do not gain any advantage from having the prompt image when attempting subject identification.

4.3 Analysing identification accuracy difference between images

Some images are better at capturing personal features, resulting in higher subject identification accuracy. Image 5 has a perfect track record for both experiments, all reconstructions from this prompt were correctly classified to the right subject. Image 10 on the other hand performed the worst for all subjects. The image prompt with their reconstructions can be viewed in figure 10. A directly noticeable difference is the background color of the different images. Image 10, which has a white background also produced a white background in all reconstructions, however image 5, which has a black background produces backgrounds of varying color in the reconstructions, making them easily distinguishable. Graphing the amount of black areas each of the reconstructed images gives us figure 11. This indeed shows that figure 5 has the most darkness followed by figures 8, 7, 3, 9 and 2. These images are also part of the images returning the highest subject accuracy, images 1, 3, 5, 7, 8 , 9. The lowest images 6 and 10 also have the lowest *mean* darkness. The only inconsistency of this link are images 1 and 2. This shows that although the darkness does not provide perfect identification, it helps maintain personally identifiable features.

Three theories have been identified to explain the role darkness can play in subject identification. Dark sections in a prompt image mean no light is getting detected by the subject's retina and thus no information is entering the visual cortex. These black sections could lead to a reduction of brain activity in certain areas as the subject is not seeing anything, thus allowing the fMRI scan to pick up the background noise of the subjects brain. This noise could be what is manifesting in the dark parts of the image and creating these varying colours. Alternatively, the colours could also be remnants of the individual strategies used by the model for reconstruction, manifesting when the fMRI presented no strong brain activity. The noise and strategy theory becomes more conceivable when looking at the reconstructed images per subject, each subject seems to have a different colour tinge present on most of their images. This however can not be found to be conclusive due to the lack of data (10 images per subjects) and the few exceptions to this rule.

The second derisive theory points out the brain's generative behaviour when confronted with a lack of information. When one closes one's eyes or looks at the wall in a dark room, they see shapes appearing and disappearing as the brain tries to make sense of what it cannot see. A similar process might be detected by the fMRI scan and remain in the reconstructed image. These generated features come purely from the subject and are individual enough to be used as personal features for better classification.

For the third theory we turn to Bannert et al [14]. In the paper they showed subjects grey scale images of objects that have a strong inherent colour association, such as a banana with yellow. When predicting the colour of the seen images from fMRI scans of the visual cortex using a network model, the model was able to predict the associated colour, even though the image prompts were gray scale. Bannert et al concluded that 'memory colour' or the colour the subject associates with images can influence the seen object as early on as in the visual cortex. Associations from memory can be very subjective when viewing complex images such as stained glass windows (image 5). This could explain why every subject perceived a different colour in the darkness of image 5, thus making it easier for the CNN to differentiate them. As no colour is present as input, 'memory colour' holds more value. This idea shows that a person's subjective memory information also remains within the reconstructed image, thus adding another layer to the privacy risk.

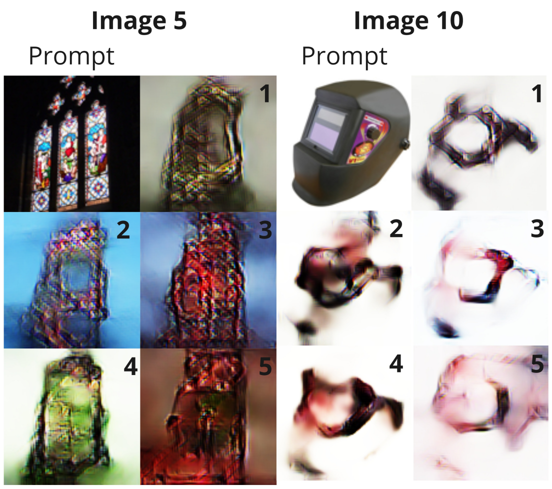


Figure 10: Image 5 with reconstructions on the left. Image 10 with reconstructions on the right

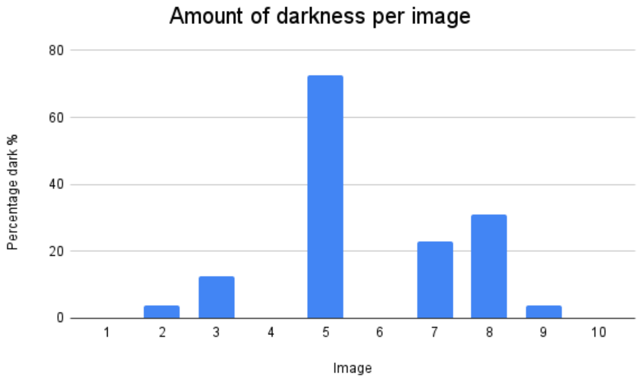


Figure 11: Amount of darkness present in the 10 different test images.

4.4 Limitations

This comes with limitations. Firstly, as the subjects are identified using a neural network model, labelled training data is necessary to get accurate classifications. Reconstructed images known to be from a subject are necessary for the presented model to re-identify that subject. This research does not explore the possibility of classifying reconstructed images to identify subjects, using a network trained on other subjects.

Secondly, the results were derived using only one of the available image reconstruction models. Further research using other models would be needed to verify that personal features also remain during image reconstruction with other models. If generative models reconstruct images without personal features they could also provide a defence against identification attacks. Zhang and al [15] mentions that generative adversarial models have already been used for de-identification of data in images. The generative process creates new samples from a probabilistic distribution, replacing private objects in the image with synthetic objects. This has already been done by Samarzija and al [16] for faces, generating new features to render them un-identifiable while keeping the face realistic.

Finally, this study was performed using a relatively small data-set of 250 images. This did not allow to research if some images return better accuracy for certain subjects, as the data-set only contained one example of each prompt image per subject, thus making any potential result insignificant.

5 Responsible Research

5.1 Ethical Research

Computer-brain interfacing is a new field that is already proving to have a variety of useful applications, such as restoring sensory and motor functions or addressing neurological disorders [2]. Brain imaging is an important step in understanding how the brain sees and thinks. However, the development of research in this field entails ethical and legal risks. They must be considered in conjunction with the development of new technologies to prevent potential breaches and unwanted consequences.

When collecting data, it is of paramount importance to ensure the subject's safety and avoid risks to their health. The brain scans used as data for this research were gathered through functional Magnetic Resonance Imaging, a non-invasive sensory technique that uses magnetic impulses to measure the oxygen content in the brain, which is then used to map brain activity. These impulses have been thoroughly researched and marked as safe [17]. They are now being used widely all over the world.

Consent is also important to consider for brain imaging. First, according to Tang and al [18], subject cooperation, which implies consent, is currently necessary for successful semantic extraction. If subjects do not pay full attention to the prompt for the entire duration of the scan, the extracted data will not be sufficient to reconstruct their brain activity.

Second, extracting information from unconsenting individuals would constitute a breach of ethical principles and of legal rules set down in particular in the EU's General Data Protection Regulation [19]. Data protection is necessary when working with highly sensitive data such as medical data. As mentioned by Ienca and al [20], cases of "brain-hacking" are already being discussed but the risks associated with the misuse of these technologies remain largely unexplored. Like with cyber-security, it is important to research the extent to which personal data can be extracted from different technologies to then warrant precautions and the implementation of defences to protect them. This paper is drafted with the intention to make users of brain imaging technologies aware of the privacy risks linked to re-identification.

5.2 Transparent and reproducible research

All data and code used in this research are freely available for non-commercial use. The steps and hyper parameters used to generate these results are listed in the methodology and result sections, allowing for the results to be reproduced. It is to be noted that some randomness is used when training neural networks. Thus results may deviate by an insignificant margin. The model was run multiple times to ensure that results did not originate from an outlier.

6 Conclusions and Future Work

6.1 Conclusion

Image reconstruction entails recording the brain activity of a subject using fMRI, while they are shown an image prompt. A neural network model will then use the brain activity to reconstruct the seen prompt image. The question posed in this paper was if it is possible to identify the subject whose brain scan was used as input for the reconstruction of an image prompt.

To achieve this goal, a CNN model was proposed to classify reconstructed images to the subjects they came from. This model was shown to identify subjects from their reconstructed image with an average accuracy of 90.4%. This proves that personal features are present in the reconstructed images, thus allowing the CNN model to identify the corresponding subject. The results also show that adding the original prompt images, as further input alongside the image reconstructions, does not lead to a significant increase in performance.

Further research was conducted on the performance differences between subjects and between images. Some subjects were shown to be more easily identifiable than others. This correlated with the quality of their image reconstructions. Subjects who tend to produce higher quality image reconstructions appear to be easier to identify. Although this trend holds globally it does not always hold when the difference in quality is minute. We thus conclude that image reconstruction quality plays a role in identifying the subject as it carries clearer personal features, but it is not a determining factor.

When comparing identification accuracy over different image prompts, certain trends also emerge. Some images were shown to consistently successfully identify the correct subjects. The best performing prompt images were the ones with the largest dark areas, which when viewed as reconstructions would tend to turn into a different colour for each individual. Three theories were proposed to explain this phenomenon. The first is that the colours represent the background noise of the brain read by the fMRI when it had no stronger colour signals to record, as the area was black (meaning that no light reached the eyes). The second theory suggests that the colours are remnants of the strategy used by the neural network model during reconstruction. The model would employ different strategies when dealing with different subjects as input and thus produce different results. The last theory introduces the idea of 'Memory colour', referring to the colour a person links with particular images. This was shown by Bannert et al [14] to play a role in a person's perception of images even if the said colour is not present. This subjective colouring happens as early as in the visual cortex, the same place reconstructed images are taken from. The colour substituting the dark patches could thus be an incarnation of this subjective colouring, presenting itself when no other colour is present to mask it.

The personal features present in the reconstructed images constitute personal information individual to each subject. It is thus important to safeguard this information and take the appropriate precautions when using image reconstruction technology to prevent possible breaches of ethical principles and legal rules.

6.2 Future work

Having shown that some models can keep identifiable information in reconstructed images, further research into possible defences to prevent the extraction of personal information is warranted. A limitation of this study was the small size of the data set used. Extending this research to other available models and data-sets would help define the possible extent of such privacy breaches. Furthermore, investigating the material representation in the reconstructed images of the personal features the CNN uses to identify subjects could provide insight on how individuals see the world.

References

[1] G. Gaziv and M. Irani, "More Than Meets the Eye: Self-Supervised Depth Reconstruction From Brain Activity," *arXiv preprint arXiv:2106.05113*, 2021.

[2] Rakhimberdina, Z., Jodelet, Q., Liu, X., Murata, T., "Natural Image Reconstruction From fMRI Using Deep Learning: A Survey. Frontiers in neuroscience," *Frontiers in neuroscience*, vol. 15, no. 795488, 2021.

[3] Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B., "Seeing it all: convolutional network layers map the function of the human visual system.," *Neuroimage*, vol. 152, 184–194, 2017.

[4] G. G. H. A. S. F. G. T. . I. M. Beliy, R., "Advances in neural information processing systems," vol. 32, 2019.

[5] M. K. H. T. K. Y. Shen G, Dwivedi K, "End-to-End Deep Image Reconstruction From Human Brain Activity.," *Front Comput Neurosci*, 2019.

[6] T. Fang, Y. Qi, and G. Pan, "Reconstructing perceptive images from brain activity by shape-semantic gan," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 13038–13048, Curran Associates, Inc., 2020.

[7] L. R. M. Mozafari and R. VanRullen, "Reconstructing Natural Scenes from fMRI Patterns using BigBiGAN," *International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020*, pp. 1–8, 2020.

[8] A. L. G. Y. v. G. M. Seeliger K, Güçlü U, "Generative adversarial networks for reconstructing natural images from brain activity.," *Neuroimage*, vol. 181, pp. 775–785, 2018.

[9] W. L. Z. C. T. L. Y. B. Qiao K, Chen J, "BigGAN-based Bayesian Reconstruction of Natural Images from Human Brain Activity," *Neuroscience*, vol. 444, pp. 92–105, 2020.

[10] G. Gaziv, R. Beliy, N. Granot, A. Hoogi, F. Strappini, T. Golan, and M. Irani, "Self-Supervised Natural Image Reconstruction and Large-Scale Semantic Classification from Brain Activity," *NeuroImage*, 2022.

[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018.

[12] T. Horikawa and Y. Kamitani, ""generic object decoding (fmri on imagenet)"," 2018.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[14] A. B. Michael M. Bannert, "Decoding the yellow of a gray banana," *Current biology*, 2013.

[15] "Visual privacy attacks and defenses in deep learning: a survey.," *Artif Intell*, vol. 55, 2022.

[16] B. Samarzija and S. Ribaric, "An approach to the de-identification of faces in different poses," 05 2014.

[17] B. P. D. V. F. R. M. M. B. L. L. D. J. R. T. B. H.-P. L. . S. L. Herate, C., "The effects of repeated brain MRI on chromosomal damage," *European radiology experimental*, vol. 6(1), no. 12, 2022.

[18] L. A. J. S. e. a. Tang, J., "Semantic reconstruction of continuous language from non-invasive brain recordings.," *Nat Neurosci*, vol. 26,, p. 858–866, 2023.

[19] "General data protection regulation."

[20] H. P. Ienca, M., "Hacking the brain: brain–computer interfacing technology and the ethics of neurosecurity.," *Ethics Inf Technol*, vol. 18), p. 117–129, 2016.