Noise Attacks as a First Layer of Privacy Protection in Semantic Data Extraction From Brain Activity

Thomas Walter[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: Thomas Walter
Final project course: CSE3000 Research Project
Thesis committee: Xucong Zhang, Nergis Tömen

Abstract

This paper explores using synthetic noise superimposed on fMRI data to selectively impact the performance of the Generic Object Decoding (GOD) model developed at Kamitani Lab. The GOD model predicts image categories that subjects viewed, based on their recorded fMRI brain activity. To evaluate how selective the noise can be in impacting performance, a new measure is proposed: the Noise Specificity Score (NSS). A highly selective noise pattern would allow for protecting sensitive data while retaining performance on non-sensitive categories. An evolutionary approach of iteratively mutating noise candidates was chosen to maximise the NSS.
Scores ranging between 0.75 to 0.8 were achieved across three different categories. The results also further support the GOD hypothesis of analogous structures between large image classification algorithms and the human visual cortex.
Limitations included computational capabilities and inherent challenges of evolutionary algorithms. Consequently, multiple opportunities for future research are proposed. These include improvements to the current approach, specifically increasing population and generation sizes of the evolutionary algorithm, enabling adaptive learning rates to escape local maxima, but also another approach based on evaluating an individual voxel's impact on the different category performances. Additionally, a novel architecture is proposed for future research that leverages pre-generated noise candidates and selects the most promising one for each visual stimulus. This is predicted to achieve much better results at lower training times.

Keywords GOD, Noise Attacks, BCI, Visual Stimulus Categorisation, fMRI, Thought Privacy

1 Introduction

In the past decade, advancements in Neuroimaging and Machine Learning have resulted in increased efforts to reconstruct or categorise seen images (hereafter visual stimuli) from captured brain activity (Rakhimberdina, Jodelet, Liu, & Murata, 2021). Breakthroughs in this field would lead to an increased understanding of how the brain processes images, represents concepts, and enable new treatments for sensory conditions (Wandell, Dumoulin, & Brewer, 2007). Researchers also pursue techniques to better understand dreams (Cowen, Chun, & Kuhl, 2014).

While many researchers focus on the application and improvement of the techniques, little to none are concerned with their safety. Semantic understanding of captured brain data brings ethical risks with it. How do we ensure privacy of thought? This paper therefore proposes one approach to ensure privacy of thought in semantic data extraction from brain activity: the possibility of synthesizing noise which, added to the original brain data, will reduce performance of categorisation models on specified image categories. In the future this could serve as a first layer of protection of thought privacy while maintaining performance on non-sensitive categories.

Noise attacks on machine learning models have been readily explored, mainly on Convolutional Neural Networks (CNNs) which are prone to reacting sharply to small disturbances of their input (Goodfellow, Shlens, & Szegedy, 2014). This paper however focuses on applying a noise attack to the Generic Object Decoding (GOD) model developed at Kamitani Lab (Horikawa & Kamitani, 2017). The model is based on the hypothesis that the human visual cortex, responsible for processing images, shares analogies with large neural networks trained to categorise images. By exploiting those similarities, one can use large pre-trained categorisation models to simplify the visual stimulus categorisation. The model uses functional magnetic resonance imaging (fMRI) recordings of five subjects that are shown images from the ImageNet database.

Putting the above together, this paper aims to answer the following research question:

How accurately can synthetic noise, that is superimposed on the input data, impact the categorisation performance of the GOD model on a specific image category, without reducing performance on other categories?

Having explored the current context of noise attacks and high-level information extraction from brain data, several steps are still needed to answer this question. Section 2 details the more formalised problem description, identifies key concepts necessary to understand the problem, defines measures, and provides an account of the experimental setup and experiments performed. The consequent data and most interesting features of the results are presented in section 3. Two discussions ensue: firstly, section 4 discusses the technical implications of the results, especially in relation to the research question. Secondly, section 5 discusses the ethical implications of the results, both from a stakeholder perspective and a reproducibility perspective. Finally, we end with a summary and conclusion in section 6.

2 Problem Description and Experimental Setup

This section declares relevant terms that are needed to pose and understand the precise problem statement.

2.1 Feature Vector

To exploit analogies between the human visual cortex and large CNNs that classify images, GOD, rather than trying to predict image categories of visual stimuli directly from the brain activation data, predicts so-called feature vectors of the CNNs (Horikawa & Kamitani, 2017). A feature vector in our case, is a subset of the output of a layer of the neural network, given an image to categorise as input. It is a subset because the outputs of the layers can be high-dimensional.
The idea of using feature vectors is, that when categorising an image, the neural network is assumed to extract higher- and

higher-level information with each layer. If one can predict the neural network feature vector from the brain activity directly, one is essentially extracting that high-level information from the brain activity.

The feature vectors for every image used in the experiments were precomputed and are part of the GOD dataset.

2.2 Noise Specificity Score (NSS)

To evaluate the results of the experiments, a measure must be established which captures the impact of noise on the categorisation performance of the model on the attacked category, in relation to the performance retained in the other categories.

A first proposed measure was subtracting the performance of the attacked category from the average performance of the other categories. This, however, lead to overfitting in trial runs, as the noise significantly improved performance of some categories to maximise the NSS. A better measure was therefore developed, which aims at minimising the impact of the noise on the non-attacked categories overall:

$$NSS = 1 - p_{\text{attackedcategory}} - \frac{1}{n-1} \sum_{i \neq i_{\text{attackedcategory}}} \left| p_{i_{\text{original}}} - p_{i_{\text{noise}}} \right|$$

Where p is the respective GOD performance of a category and n is the total number of tested categories. The GOD performance measure for categorisation accuracy works as follows. First, predict the feature vector given the brain activation for a certain visual stimulus using a linear regression model for each element of the feature vector. Then compute the Pearson correlation between the predicted feature vector and a category-averaged feature vector of every image category in ImageNet. If the correlation for an arbitrary category is higher than for the correct category, the model has made a false categorisation prediction. Looking at the percentage of categories whose averaged feature vectors have a lower correlation to the prediction than the correct category, therefore, reflects the categorisation accuracy of the model for that category. Images from 50 different categories that were not part of the training data are tested in the experiment.

In later, longer runs, the difference between the two measurements was negligible, however the second measurement was kept because it more accurately reflects the research question.

2.3 Problem Statement and Proposed Solution

We are interested in finding a noise pattern $\{n_0, \ldots, n_N\}^T$ with N the number of voxels used to test the linear regression models, which maximises the NSS. Formally stated:

$$\text{Find a noise pattern } \{n_0, \ldots, n_N\}^T$$
$$\text{with } max(NSS(\{n_0, \ldots, n_N\}^T)).$$

An efficient and generalisable method to maximise any quantity is the evolutionary approach (Eiben & Smith, 2015). Here, a "population" of noise candidates is randomly generated, applied, and evaluated using the NSS. The highest performing candidate produces "offspring": mutations to the candidate are applied randomly by adding a uniformly distributed random value of a fixed range to each entry. This is essentially the same as taking random samples from the solution space and following the most promising ones.

The technique is widely applicable, as it treats the underlying model as a black box, however, it is sensitive to the initial candidates and to local maxima. Part of those drawbacks can be mitigated by increasing the population size and number of generations.

Experiments were performed on three initially high-performing categories (categories 20, 33, and 42) with 12 generations and a population of 20.

The mutation range and initial noise depth were determined using trial runs, as well as an evaluation of the overall performance after applying noise of varying depths. A depth of 2 and a mutation range of 3 gave promising results.

2.4 Lightweight GOD: Adjustments to the original model

The original experiments were run on many different configurations requiring the processing of large amounts of data. To keep computation intensity low while retaining the highest quality of results, the experiment was run for only one subject as performance varied little between subjects, and with the Region of Interest in the cortical area (ROI) and neural network layer that ranked among the highest prediction accuracies: This combination is using the higher-level visual cortex ROI V4 (with 500 voxels) to predict features of units in the higher-level layer CNN6. Intuitively it makes sense to choose higher-level ROIs and layers, as categorisation happens on the higher levels. Using only this combination, the categorisation accuracy of the model is similar to that when using all ROIs and layers.

Further reductions that largely retained categorisation accuracy were made. Predicting 100 CNN6 layer units instead of 1000 when using the regression models worked well in combination with using linear regression instead of sparse linear regression.

Lastly, the original paper distinguishes between two types of experiments: perceived and imagined stimuli. The experiments focused on the former, as the initial performance on imagined stimuli was relatively low, to begin with. In some places, however, it is interesting to see how the noise synthesized for the perceived type impacts the performance of the imagined type.

3 Results

In the following, the result for running the evolutionary algorithm for 12 generations, with a population of 20 noise candidates, for each of the three categories will be displayed and described.

3.1 Category 20

The training curve for the attack on category 20 is very steep, jumping to an NSS of 0.72 after the first generation, with only minor improvements in the subsequent generations (See figure 2) and a final NSS of 0.8. The categorisation performance overall starts out high, around 0.7, then drops and slowly starts to rise again (see Figure 1). When comparing the performance of the final result to the performance without noise

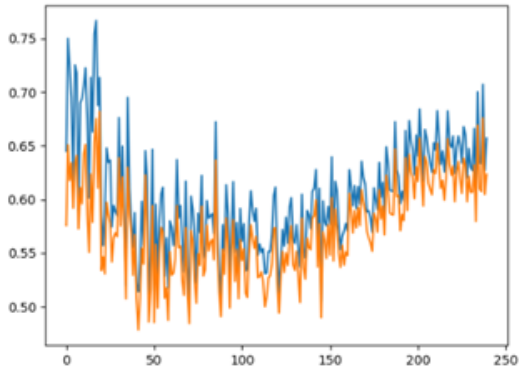

Figure 1: Overall Categorisation performance of each noise candidate during training of the attack on category twenty. Blue: Perception type, Orange: Imagined type

(figures 3 and 4) of the individual categories, one can see a deep spike at category 20, with most other categories being slightly affected. Spikes also occur at categories 23 and 15. Performance for category 46 was improved by a large margin.

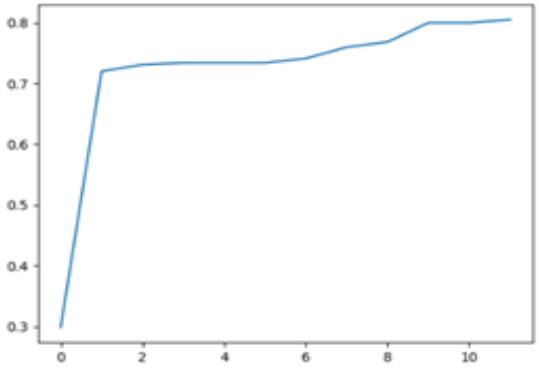

Figure 2: Maximum NSS score for each generation during training for category twenty.

3.2 Category 33

The training curve for the attack on category 33 is less steep than for category 20, however diminishing returns can be observed towards the end, with a final NSS of 0.74 (see Figure 6). The category identification performance, similarly, to category 20, starts out high, then drops and slowly recovers (see Figure 5). There is a clear spike at category 33, its performance almost reduced to 0, with multiple other categories spiking (see figures 7 and 8). Most categories received less impact however, indicated by the final NSS. Categories 12 and 46 received performance improvements.

3.3 Category 42

For category 42, there is a steep learning curve, with a minor breakthrough in the last generation, resulting in a final NSS of 0.78. The categorisation performance during training looks like that of the other categories, in that there is a brief spike

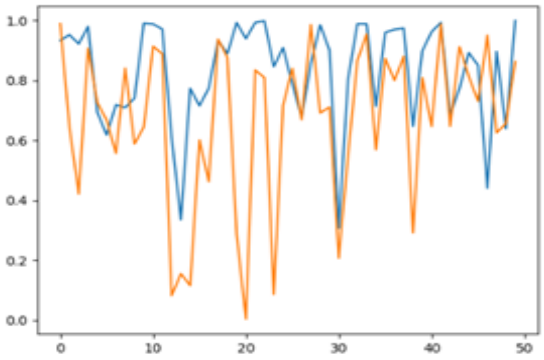

Figure 3: Categorisation Performance across all fifty test categories without (blue) and with (orange) added Noise, Category Twenty.

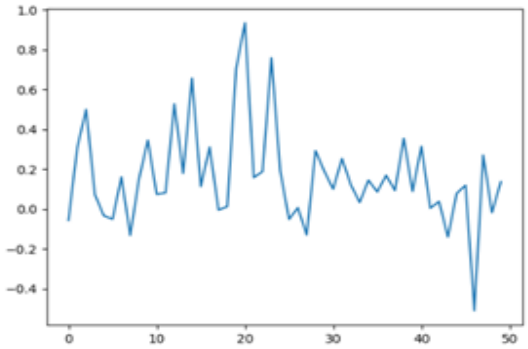

Figure 4: Difference between Categorisation Performance without and with added noise, Category Twenty

in the beginning, followed by slow recovery. Here, however, a flat or even dwindling trend ensues. The performance for category 42 nearly reached 0. The same is true for category 6. Several other categories were impacted severely, while most categories stayed around a performance reduction of 0.2 to -0.2. Category 46 received strong performance increases once more.

4 Discussion of the Results and Limitations

This section puts the results into context, attempts to answer the research question and discusses improvements, limitations and further research that is opened from this work.

4.1 Explaining the Results

Overall, the results are promising. In each case the performance of the targeted category was reduced to almost zero. Furthermore, most other categories were left impacted at a much lower rate in every case.

To answer the research question, it is possible to selectively reduce performance of arbitrary single categories in the GOD model, while keeping other categories performant.

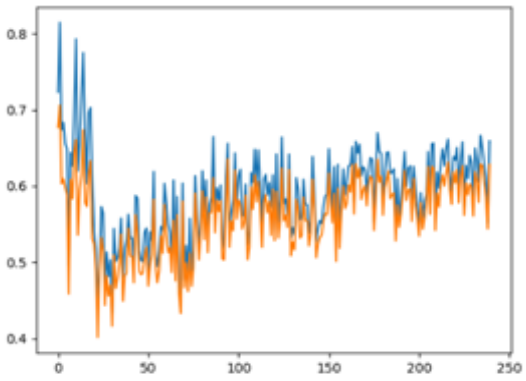

Figure 5: Overall Categorisation performance of each noise candidate during training of the attack on category thirty-three. Blue: Perception type, Orange: Imagined type

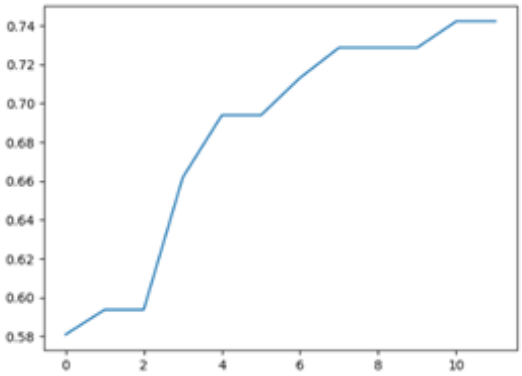

Figure 6: Maximum NSS score for each generation during training for category thirty-three.

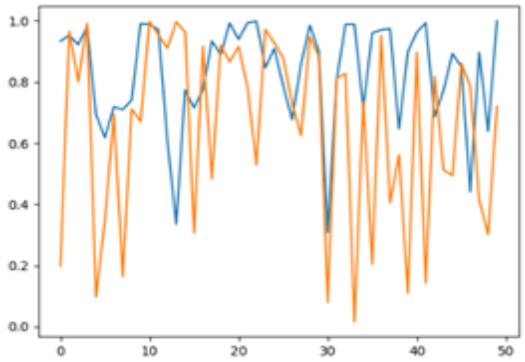

Figure 7: Categorisation Performance across all fifty test categories without (blue) and with (orange) added Noise, Category Thirty-Three

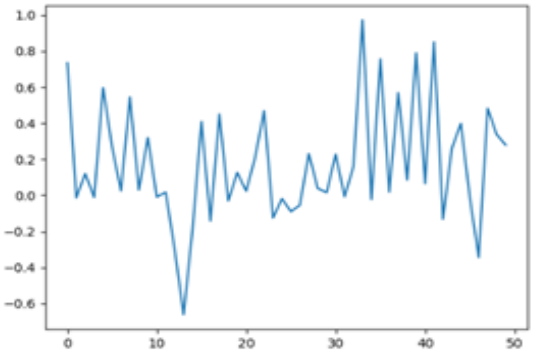

Figure 8: Difference between Categorisation Performance without and with added noise, Category Thirty-Three

While some other categories were impacted as well, the NSS scores of 0.75-0.8 in combination with the highly reduced attacked category performances indicate a comparably low average performance change of around 0.2. It is therefore viable to consider noise sources a possible physical layer of security and privacy protection in this type of brain interfacing. The physical implementation of such a system, however, is the subject of future research.

The steep learning curves in conjunction with the performance spikes at the beginning of training are a direct result of the proposed NSS measure. It is easy to find noise candidates that decrease the performance of the attacked category, as any noise that decreases most categories' performances suffices. This results in an increase in the NSS and a decrease in overall performance. Only when the performance of the attacked category is close to zero does the selectivity part of the measure come in: the overall performance starts rising slowly, as noise candidates are found which are more selective. This also slowly raises the NSS, resulting in the depicted training curves.

4.2 Improvements

The results can be improved upon in two separate ways. Firstly, breakthroughs as seen in the attack on category 42 indicate that increasing the number of generations, while showing diminishing returns, could further improve the NSS score. Secondly, the lack of progress in that attack between generations 2-10 indicates a lack of variety for any evolutionary process to show effect. Extensions to the evolutionary algorithm, such as selecting the top k performing candidates as parents for the next generation instead of only the top one candidate, or dynamically adapting the mutation range when progress is stagnating, require future investigation (Eiben & Smith, Evolution Strategies, 2015). These improvements would also counter the limitation of being sensitive to the initial noise candidates and getting stuck in local maxima.

Different approaches to the maximisation problem should also be considered, for example, an approach inspired by gradient-based noise attacks (Goodfellow, Shlens, & Szegedy, 2014). Such an approach investigates the influence of each voxel on the categorisation performance. This would also provide answers to localisation questions. A high localisation of brain functions (in terms of processing concepts and categories) would imply promising performance with this approach.

While the original evolutionary approach operates somewhat blindly, both of these are attempts at guiding the random exploration of the solution space of the noise patterns, to reduce

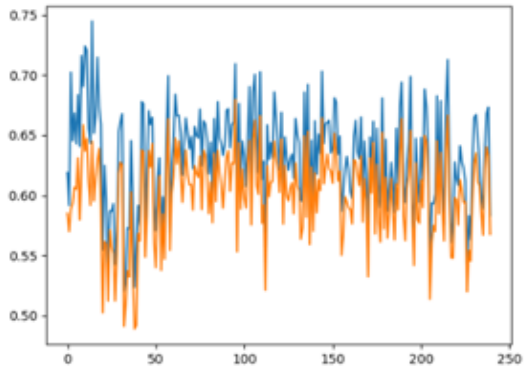

Figure 9: Overall Categorisation performance of each noise candidate during training of the attack on category forty-two. Blue: Perception type, Orange: Imagined type

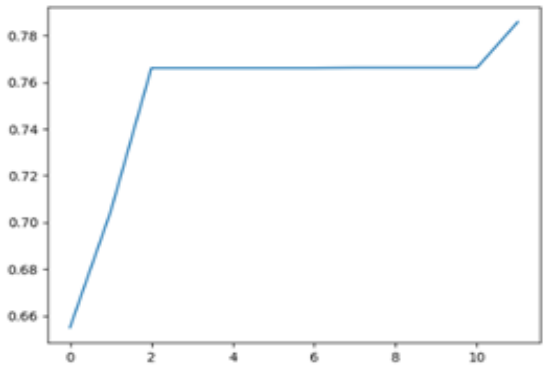

Figure 10: Maximum NSS score for each generation during training for category forty-two.

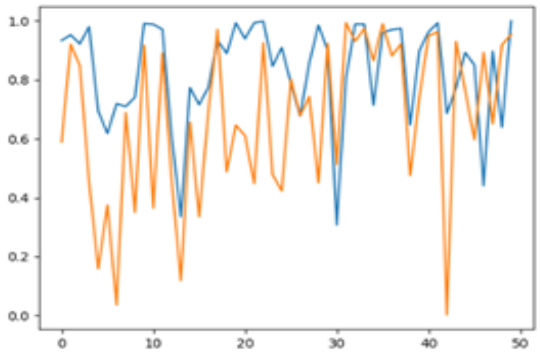

Figure 11: Categorisation Performance across all fifty test categories without (blue) and with (orange) added Noise, Category Forty-Two

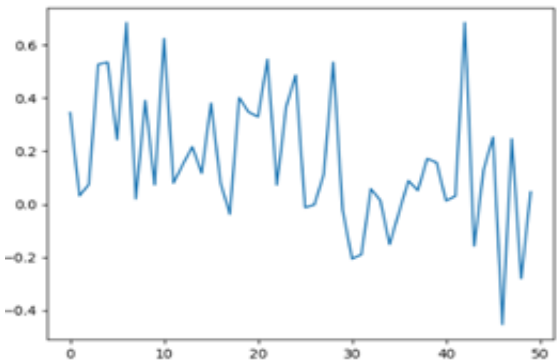

Figure 12: Difference between Categorisation Performance without and with added noise, Category Forty-Two

the computational intensity while maximising results.

4.3 Answers? No, more questions!

While this work answered the initial research question, it also brought with it many more.

It is noteworthy for example that, even though the performance of the imagined type experiments was not considered during training, the performance graphs show that their performance follows that of the perceived type experiments very closely. It is to be expected that the attack therefore works for that category as well, which should be investigated further in the future.

It could be interesting to investigate which categories are most likely to drop in performance together. In the original GOD, semantic distances between categories, computed from graphs resulting from the ImageNet categories, played a big role when analysing performance (Horikawa & Kamitani, 2017), supporting the hypothesis that categories that are closely related, are also spatially close in the brain, which would make it harder to maintain selectivity. While the sample size in this work is too low to draw meaningful conclusions, trial runs did give the impression that some categories were more likely to drop with other categories. A future investigation should explore whether this is indeed true, and if so, whether it is due to initial noise conditions or whether it is encoded in the anatomy of image processing both in the brain and in neural networks.

More compute power should also be employed to investigate more attacked categories and contingencies between them and the many other configurations of the initial GOD experiments (Horikawa & Kamitani, 2017).

Lastly, an initial mistake in the approach lead to an entirely new idea to approach the problem. In initial versions of the project, the linear regressors saw a different noise pattern for each image when predicting the feature vectors. The results were much faster increases in NSS at much lower generation and population sizes. This was an accident, as the mechanism investigated should employ a static noise source to make physical implementations easier. A resulting architecture for a better performing noise attack, however, is possible and shall be proposed here: assuming an adaptive noise source exists, a system selects the best noise from an array of precomputed noise patterns, where each noise pattern was generated for a specific noise category. Such a system could be trained separately and yield much better results.

5 Reproducibility and Integrity

Reflections on the ethics of this research project in terms of societal stakeholders ensue, followed by a discussion of the research integrity.

5.1 Moral Ethics

While posing exciting possibilities in the future, brain-computer interfaces (BCIs) are also a huge ethical risk. There are three main problems that need to be addressed before the technology can be released to the public.

Firstly, there is the discussion of transhumanism. At its core, the debate revolves around whether humans should strive to enhance their bodies technologically, or if that essentially brings us too far away from being human. This debate is difficult to judge as it highly depends on what is defined as being human. Early applications of BCI's circumvent this issue by focusing on enabling disadvantaged people to lead more normal lives again. However, once the technology is ready, there is no doubt that some company will bring it to the masses.

Secondly, and most specifically for this topic, there is the problem of the most fundamental private space that we as humans possess. This space is our thoughts; for most of human history, it was assumed to forever be private. Freedom of thought is one of our most basic human rights, and it is one that defines us to a deep degree. Any technology that is capable of extracting information from the brain, especially non-invasively and without extortion, can have far reaching consequences for this right. Adversaries may utilise such devices to access restricted information, to violate the most intimate privacy sphere a person has and wield a tool of total oppression of thought.

Lastly, from a different perspective, a countermeasure to unwanted brain-reading could also always be misused to fabricate or tamper with existing brain-data, to carry out adversarial attacks. A person could for example be falsely accused of thinking a certain thing, even though the attacker only tampered with the data in a way that makes it look like they did. This is because with the ability to tamper with data to obscure it, comes the ability to tamper with data to plant false evidence. These lines of attack should be further visited in future works.

It is ever more necessary to create theoretically secure safeguards against misuse of the technology, to always put agency over a person's thought fully into their hands. It should not be physically possible to misuse the technology, because otherwise, someone will do it eventually.

5.2 Reproducibility and Integrity

It is of utmost importance to approach any research project with integrity to maintain the safety of all stakeholders. In this case, any human subjects have given consent to their brain activity being recorded. Societal stakeholders have been mentioned above, and the research only involved the author, their laptop, the work of Kamitani Lab and in some instances the supervisor and project colleagues during meetings.
The research performed can be reproduced by cloning the code repository from GitLab[1], downloading the subject data as specified in the README file and then running the "analysis_FeaturePrediction.py" and "gradientnoiseattack.py" files (the latter unfortunately named, as it does not use gradients). There is no need to download the actual training images. Any design choices are laid out in the sections above. All generated noise is based on a seed, so the results are deterministic.

6 Conclusion

We have shown that it is possible to synthesize noise that selectively impacts the categorisation performance of specified image categories while retaining the performance of other categories when using the Generic Object Decoding model developed at Kamitani Lab to extract semantic data from brain activity recordings of human subjects looking at images.
The chosen approach is an evolutionary algorithm that selects the top-performing noise candidate out of an initial set and mutates it to create a new generation repeatedly.
To rank the noise candidates, a new measure, the Noise Specificity Score (NSS), is proposed, which balances the performance of the non-targeted categories with the non-performance of the targeted category. NSSs of around 0.75 to 0.8 (maximum achievable NSS is 1) were reached on three separately targeted image categories.
While providing a viable tool for privacy protection in semantic data extraction systems, compute restrictions and common pitfalls of evolutionary algorithms limited the results. Several different approaches for future works are therefore proposed. Improvements to the original model, such as increasing population and generation numbers, and making the algorithm more adaptive to different training performances, but also a new approach based on explaining the contributions of each voxel to the categorisation performance were postulated. A novel architecture based on a system that selects the best noise from a set of pre-generated candidates for every visual stimulus was also proposed, motivated by faster and better results.
It is to be expected that the accuracy, speed, and portability of brain data extraction systems will only increase in the near future. With exciting possibilities on the horizon, we cannot forget about keeping our innermost thoughts safe.

Bibliography

Cowen, A. S., Chun, M. M., & Kuhl, B. A. (2014). Neural portraits of perception: Reconstructing face images from evoked brain activity. *NeuroImage*.

Eiben, A. E., & Smith, J. E. (2015). Evolution Strategies. In A. E. Eiben, & J. E. Smith, *Introduction to Evolutionary Computing* (p. 101). Berlin, Heidelberg: Springer.

Eiben, A. E., & Smith, J. E. (2015). What is an Evolutionary Algorithm? In A. E. Eiben, & J. E. Smith, *Introduction*

[1] https://gitlab.com/ruleroftthedarkrealm/noise-attack-privacy-for-brain-information-extraction/-/commits/ResearchProjectThomasWalter

to Evolutionary Computing (p. 25). Berlin, Heidelberg: Springer.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *arXiv preprint*.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*.

Rakhimberdina, Z., Jodelet, Q., Liu, X., & Murata, T. (2021). Natural Image Reconstruction From fMRI Using Deep Learning: A Survey. *Front, Neurosci.*

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual Field Maps in Human Cortex. *Neuron*, 366-383.