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

Selectively impacting the performance of a machine learning model that extracts information from brain activity, after training and on arbitrary categories, to ensure privacy in brain interface applications

Thomas Walter, T.C.Walter@student.tudelft.nl

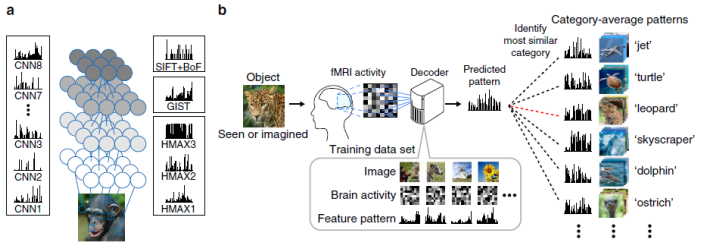
How accurately can synthetic noise, that is superimposed on the input data, impact the categorisation performance of the GOD model on a specific image category, without reducing performance on other categories?

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

Specificity evaluation metric: how selective is the noise?

Goal: Balance Impact on performance on attacked category with impact on performance of the other categories.

p is the GOD categorisation performance of the respective category, n is the number of categories

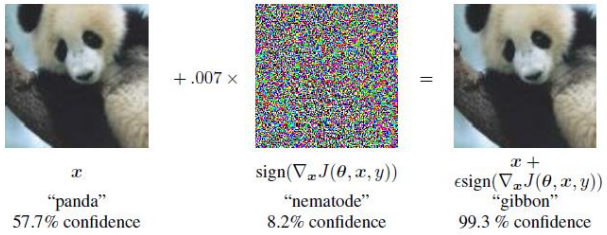


Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*.

Generic Object Decoding (GOD)

Goal: Categorise what object a person is looking at, based on their recorded brain activation

1. Compute the outputs of various layers (called feature vectors or feature maps) of neural networks for many different image categories
2. Train linear regression models to predict the output of the layer from a human's brain activations who is looking at the same images
3. Show the model a novel example and predict the neural network output
4. Compute correlation of prediction to category-averaged neural network outputs of all ImageNet categories
5. Evaluation: percentage of categories which have a lower correlation with the prediction than the correct category

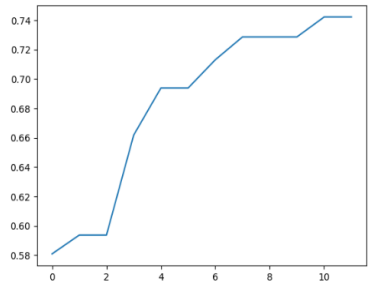


Noise attacks

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

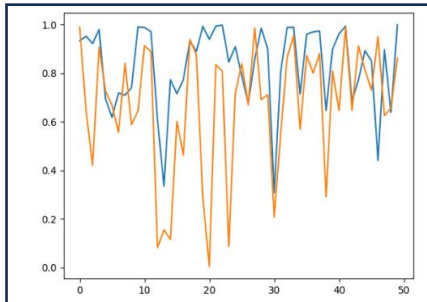
Goal: add (imperceptible) perturbations to the input of machine learning classifiers to manipulate their predictions

- Gradient-based approach: compute the loss gradients of the neural network in the direction of the desired output
- → difficult here, as we have an array of linear regressors, each predicting a different entry of the feature vector, and the targeted performance
- Chosen approach:
  an evolutionary algorithm that iteratively generates better and better noise candidates (12 generations with a population of 20, top performer selected as parent for next generation)
- Add the noise in Step 3



Maximum NSS score for each generation during training for category 33 (0.74 achieved)

- A steep increase at first indicating the algorithm exploiting the effect of general noise on all categories, including the attacked category
- Flattens out once attacked category performance is close to 0, starts lifting performance of other categories



Performance across all 50 categories before (blue) and after (orange) the noise attack on category 20

- Performance of attacked category reduced to almost 0
- Most other categories follow the original performance closely, with some exceptions
- Potential semantic connections or localisation effects between the commonly affected categories should be investigated in future works

Limitations and future work

- Sensitivity to local maxima and initial conditions -> increase population and generation sizes, and adaptive mutation rates
- Consider approach based on individual voxel contributions to each category performance
- Consider approach based on selecting best noise candidate from a set of precomputed ones for each image