Master Research Proposal Nr:

Lukas Böhm November 30, 2023

Topic: Developing New Approaches For Measuring Invariances Between Models Using Model Metamers

Deep Neural networks have proven themselves to be reliable models of sensory systems. Even without custom-engineered architectures, their hierarchical structure allows them to learn task-relevant representations at a high level and even out-perform humans when properly trained. This indicates that humans and machines share invariances [1]. However, their complexity also reveals issues and attack surfaces. Bad performance when using out-of-distribution data and unpredictable behavior with adversarial examples are just some of the flaws.

To get a better understanding of what invariances are shared between humans and machines, as well as finding the stage where they diverge, we can use **model metamers** [1]. We define metamers as a pair of stimuli that are physically distinct, but elicit the same response in a model. Previous work attempted to investigate what invariances are shared between humans and models by the "metamer test". Humans are presented with model metamers and have to decide which class the image belongs to [2]. The model metamers are regarded as "passing" if the human can correctly determine the class the reference stimuli (from which the metamer was generated) belongs to. This version of the metamer test does not quite capture the essence of model metamers. It just compares the output of a linear classifier (which was added to every tested model). Although comparing the actual brain activation is impossible the paper surprisingly found that feeding model metamers into other models instead of showing them to humans, behave almost identically to humans (94% correlation) [1]. Therefore, it is sufficient to just use other models to evaluate model metamers instead of relying on humans. This opens the doors for new approaches to generating and evaluating model metamers, as will be described below.

In this thesis we aim at advancing the field of model metamers, with particular emphasis on three aspects: model metamers generation, model metamers evaluation, and transferability of model metamers.

- **Model metamers generation.** Counterimpose to literature in human metamers generation [3, 4, 5], most model metamers genration approaches rely on an iterative optimization of random noise using gradient descent. This procedure is inefficient and, in the best cases, it takes about 20K iterations before a successful metamer is generated. To address this issue we will focus in improving generation performance for model metamers involves utilizing various optimizers, such as momentum-based ones, and modify the optimization procedure starting values closer to reference stimuli rather than random noise. These approaches aim to reduce iteration requirements, minimize the risk of local minima closer to noise. Additionally, we will test the effectiveness of optimizing N models simultaneously, to generate more robust model metamers that are subject to less idiosyncratic invariances [1].
- **Model metamers evaluation.** Previous work demonstrated that it is sufficient to test metamerism on a reference set of (artificial) models, since it correlates with humans behavioral reponse. However, the testing framework in state-of-the-art studies relies on recognizability metrics, specifically classifier accuracy, which may neglect the internal activations and fail to measure the degree of metameric similarity. To address this, a proposed "representation test" explores statistically relevant similarities in internal representations across models. More specifically, we will evaluate whether model metamers still preserve similarity in their respective representations that are statistically relevant (e.g., Pearson correlation) for a reference set of models. Experimentally, some visualization approaches (like [6, 7]) could be used for comparing the representations of model metamers in different models at different stages.
- **Transferability of model metamers.** By employing a diverse set of proven architectures, both standard and robust, from model zoos like PyTorch and RobustBench, the transferability test will involve generating model metamers on one model and assessing their performance on a subset of different models. We intend to compare different architectures and training styles.

The proposed work comprises the following key components:

- Literature research with a special focus on metamer generation and evaluation strategies.
- Improve metamer generation by assessing the relevance of starting values (noise vs. original input), optimization algorithm (gradient-based vs. momentum), and by exploring a distributed approach (optimize N models simoultaneously).
- Develop and test a novel evaluation framework for model metamers that do not only rely on recognizability tests, but take into account the distribution of model metamers representations (i.e., Pearson correlation on distances between representation - more details given above).
- Investigate transferability of model metamers:
  - Are model metamers generated by one model also model metamers for other architectures?
  - Are model metamers generated by one architecture also metameric to differently trained instances of the same architecture?

This comprehensive analysis will shed light on the effectiveness and requirements for building models that have shared invariances.

Project Information

Supervisors:	Dr. Dario Zanca, Dr. Christoffer Loeffler, Jonas Mueller (M.Sc.), and Prof. Dr. Björn Eskofier
Student:	Lukas Böhm
Start – End:	Dec. 1st 2023 - May 31st 2024

References

[1] Jenelle Feather, Guillaume Leclerc, Aleksander Mądry, and Josh H. McDermott. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, October 2023.

[2] Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. 2019.

[3] Arturo Deza, Yi-Chia Chen, Bria Long, and Talia Konkle. Accelerated Texforms: Alternative Methods for Generating Unrecognizable Object Images with Preserved Mid-Level Features. *2019 Conference on Cognitive Computational Neuroscience*, 2019. Conference Name: 2019 Conference on Cognitive Computational Neuroscience Place: Berlin, Germany Publisher: Cognitive Computational Neuroscience.

[4] Arturo Deza, Aditya Jonnalagadda, and Miguel Eckstein. Towards Metamerism via Foveated Style Transfer, December 2018. arXiv:1705.10041 [cs].

[5] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, April 2015. arXiv:1409.1556 [cs].

[6] Aravindh Mahendran and Andrea Vedaldi. Understanding Deep Image Representations by Inverting Them, November 2014. arXiv:1412.0035 [cs].

[7] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *ArXiv*, October 2016.