CMPE 58I: SP.TP.HUMAN-INSPIRED MACHINE INTELLIGENCE Fall 2025 Project Proposal: Impact of Anti-Biomimetic Training on CNN Emergent Behaviors

Kutay Eroğlu, 2024700051 (kutay.eroglu@bogazici.edu.tr) 28 Oct 2025

1 Introduction

This document outlines a proposed methodology to investigate the behavioral and organizational impact of a novel ”anti-biomimetic” training regimen on a convolutional neural network (CNN). The experiment is designed as a direct comparison to the ”standard” and ”biomimetic” regimens analyzed by [1]. The core objective is to determine if reversing the developmental progression of sensory input—transitioning from high-quality to degraded inputs—induces ”visual forgetting,” an amplified texture bias, or other novel emergent properties in the network's representations.

2 Network Architecture and Dataset

Model: The modified AlexNet architecture, as specified in the reference study [1], will be primarily utilized. This version features 48 first-layer convolutional filters (receptive fields, or RFs), each 22×22 pixels, a modification intended to facilitate more precise frequency-based analysis.

Dataset: All models will be trained on the ImageNet database [2].

Training Parameters: To ensure a fair comparison, the reference study's [1] parameters will be followed:

- **Optimizer**: Stochastic Gradient Descent (SGD) with a Nesterov momentum of 0.9.
- **Loss Function**: Categorical cross-entropy.
- **Learning Rate**: Constant 0.001.
- **Epochs**: A total of 200 epochs for the AlexNet architecture.

3 Experimental Training Regimens

Three distinct training regimens will be implemented from scratch:

Standard Regimen (Control): As the non-developmental control, the network will be trained on high-resolution, full-color images for the entire 200-epoch duration.

Biomimetic Regimen (Replication): To replicate the reference findings, this network will be trained on blurry (Gaussian blur, sigma=4) and achromatic images for the first 100 epochs (first half), then transition to high-resolution, full-color images for the final 100 epochs (second half).

Anti-Biomimetic Regimen (Novel): This is the core of the proposal. The developmental time course will be reversed. The network will be trained on high-resolution, full-color images for the first 100 epochs, followed by blurry, achromatic images for the final 100 epochs.

4 Analysis and Evaluation

Upon completion of training, the resulting models from all three regimens will be subjected to the following analyses:

4.1 Shape vs. Texture Bias

The exact methodology detailed in [3], as used by [1], will be followed. The 1,280 test images provided by [3], which feature explicit shape-texture conflicts (e.g., the shape of an airplane with the texture of a cat), will be used for the analysis. The percentage of classification decisions that are shape-consistent (classifying the image as ”airplane”) versus texture-consistent (classifying it as ”cat”) will be quantified.

4.2 Receptive Field (RF) Characterization

The 48 first-layer RFs will be extracted from each of the three trained AlexNet models. The spatial frequency and color sensitivity of each RF will be quantified using the specific ”Color metric” and ”Spatial frequency metric” calculations described in the supplementary methods of [1]. Scatter plots of the joint frequency and color coding for each regimen will be created to visually inspect for the emergence of distinct RF clusters, such as the ”magnocellular-like” RFs (low spatial frequency, low color) identified in the biomimetic model.

4.3 Causal Analysis: Ablation Study

To determine the causal drivers of any observed shape or texture bias, a systematic ablation study will be performed on all three models. RFs will be selectively ablated (i.e., ”zeroed out”) based on their properties. Specifically, the effect of ablating the least color-tuned (magno-like) RFs versus ablating the most color-tuned (parvo-like) RFs will be compared. The impact of these ablations on the network's shape-texture bias classification performance will be measured.

5 Hypothesized Outcomes

Standard Model: It is expected to replicate previous findings, showing a bias toward local texture over global shape.

Biomimetic Model: It is expected to replicate the reference study's findings: a markedly stronger, more human-like shape bias, driven by a distinct cluster of magnocellular-like RFs.

Anti-Biomimetic Model: It is hypothesized that this model will show a significantly stronger texture bias than even the standard model.

References

[1] Marin Vogelsang, Lukas Vogelsang, Gordon Pipa, Sidney Diamond, and Pawan Sinha. Impact of a biomimetic training regimen based on early visual experience on neural network organization and behavior. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.