

Bridging Dynamic Field Theory and Deep Neural Networks: Human Visual Attention in Naturalistic Environments

Raul Alexander Grieben

Abstract

Attention is the cognitive mechanism that selects relevant information from a continuously changing stream of sensory input. Visual attention plays a central role in higher-level cognition, yet core questions about its neural mechanisms and the generalization of laboratory findings to real-world settings remain unresolved.

Existing theories have successfully accounted for a wide range of findings, yet they continue to face challenges from enduring empirical results, while computational models often rely on task-specific assumptions. Although attention mechanisms are central in modern deep learning models, they lack capacity-limited spatiotemporal dynamics crucial for cognitive-level modeling.

This dissertation presents a neurally grounded theory of visual attention that generalizes across various laboratory tasks, including enduring empirical challenges and extends to naturalistic environments. It introduces three mechanistic neural process models based on Dynamic Field Theory (DFT), a neurally plausible framework for modeling population-level neural dynamics, with the third integrating a convolutional neural network (CNN) for feature extraction.

The first model shows how working memory representations of scenes formed during exploratory viewing guide subsequent visual search, supported by two behavioral experiments that reveal a novel scene preview effect on search efficiency.

The second model explains enduring empirical challenges, such as attention capture vs. suppression and triple conjunction efficiency, through three new principles: the nonlinear “ $N-1$ ” rule for top-down guidance, sigmoidal compression of salience, and distinct decay profiles underlying failure-prone neural competition. Using a single parameter set, it predicts 27 conditions across twelve canonical experiments (1992–2022) with high quantitative accuracy ($\text{NRMSE} \leq 8.8\%$, $R^2 \geq 0.89$, and $r \geq 0.94$).

The third model extends the theory to naturalistic settings via a hybrid DFT-CNN architecture, supporting continual few-shot online learning, autonomous visual exploration, feature-based and categorical visual search aided by scene context and memory, as well as overt and covert attention.

Together, these models provide a neurally grounded framework that addresses open questions, reconciles enduring empirical challenges, and bridges the gap between cognitive science and artificial intelligence.

Bridging Dynamic Field Theory and Deep Neural Networks: Human Visual Attention in Naturalistic Environments

Raul Alexander Grieben

Kurzfassung

Aufmerksamkeit ist der kognitive Mechanismus, der relevante Informationen aus einem sich ständig verändernden Strom aus Sinneseindrücken selektiert. Die visuelle Aufmerksamkeit spielt eine zentrale Rolle in der höheren Kognition, trotzdem sind zentrale Fragen zu ihren neuronalen Mechanismen und zur Generalisierung von Laborergebnissen auf die reale Welt nach wie vor ungelöst.

Bestehende Theorien haben erfolgreich Erklärungen zu einer Vielzahl von empirischen Ergebnissen geliefert, werden jedoch weiterhin von anhaltenden empirischen Ergebnissen herausgefordert, während Computermodelle oft auf aufgabenspezifischen Annahmen beruhen. Obwohl Aufmerksamkeitsmechanismen in modernen Deep-Learning-Modellen eine zentrale Rolle spielen, fehlt ihnen die kapazitätsbegrenzte raum-zeitliche Dynamik, die für die Modellierung auf kognitiver Ebene nötig ist.

In dieser Dissertation wird eine neuronale Theorie der visuellen Aufmerksamkeit vorgestellt, die über verschiedene Laboraufgaben generalisiert, einschließlich anhaltender empirischer Herausforderungen, und sich auf natürliche Umgebungen erweitern lässt. Es werden drei mechanistische neuronale Prozessmodelle vorgestellt, die auf Dynamic Field Theory (DFT) basieren, eine neuronal plausible Theorie für die Modellierung neuronaler Dynamiken auf Populationsebene, wobei das dritte Modell ein Convolutional Neural Network (CNN) zur Merkmalsextraktion integriert.

Das erste Modell zeigt, wie Repräsentationen von Szenen im Arbeitsgedächtnis, die während des explorativen Betrachtens gebildet werden, die anschließende visuelle Suche lenken, unterstützt durch zwei Verhaltensexperimente, die einen neuartigen Effekt der Szenenvorschau auf die Sucheffizienz zeigen.

Das zweite Modell erklärt anhaltende empirische Herausforderungen, wie z.B. Aufmerksamkeitserfassung vs. Unterdrückung und die Effizienz von Konjunktionssuche mit drei Merkmalen, durch drei neue Prinzipien: die nichtlineare “ $N - 1$ ”-Regel für Top-Down-Lenkung, sigmoidale Kompression von Salienz und unterschiedliche Selektionsprofile, die dem fehleranfälligen neuronalen Wettbewerb zugrunde liegen. Unter Verwendung eines einzigen Parametersatzes sagt es 27 Bedingungen aus zwölf kanonischen Experimenten (1992-2022) mit hoher quantitativer Genauigkeit voraus ($\text{NRMSE} \leq 8.8\%$, $R^2 \geq 0.89$ und $r \geq 0.94$).

Das dritte Modell erweitert die Theorie auf naturalistische Umgebungen durch eine hybride DFT-CNN-Architektur, die kontinuierliches Few-Shot-Online-Lernen, autonome visuelle Exploration, merkmals- und kategorienbasierte visuelle Suche unter Verwendung von Szenenkontext und Gedächtnis sowie overte und covert Aufmerksamkeit unterstützt.

Zusammen bieten diese Modelle eine neuronale Theorie, die offene Fragen beantwortet, anhaltende empirische Herausforderungen bewältigt und die Kluft zwischen Kognitionswissenschaft und künstlicher Intelligenz überbrückt.