COGNITIVE NEURAL ARCHITECTURE SEARCH REVEALS HIERARCHICAL ENTAILMENT

Lukas Kuhn[*], **Sari Saba-Sadiya**[*,†], **Gemma Roig**
Goethe University Frankfurt, Department of Computer Science, Frankfurt, Germany
The Hessian Center for Artificial Intelligence (hessian.AI), Darmstadt, Germany

ABSTRACT

Recent research has suggested that the primate brain is more shallow than previously thought, challenging the traditionally assumed hierarchical structure of the ventral visual pathway. Here, we demonstrate that optimizing convolutional network architectures for brain-alignment via evolutionary neural architecture search results in models with clear representational hierarchies. Despite having random weights, the identified models achieve brain-alignment scores surpassing even those of pretrained visual classification models - as measured by both linear encoding and representational similarity analysis. Furthermore, architectures optimized for alignment with late ventral regions perform at the level, or better, than state-of-the art models when trained on image classification tasks. These findings suggest that hierarchical structure is a fundamental mechanism of primate visual processing. Finally, this work demonstrates the potential of neural architecture search as a framework for computational cognitive neuroscience research that could reduce the field's reliance on manually designed convolutional networks.

1 INTRODUCTION

Throughout the last decade, Convolutional Neural Networks (CNNs) have emerged as powerful cognitive models capable of providing valuable insight into the neural mechanisms underlying primate visual processing (Yamins et al., 2014; St-Yves et al., 2023; Guo et al., 2025). In their seminal work, Yamins et al. (2014) demonstrated that CNNs trained to perform image classification can be used to predict brain activity with greater accuracy than previously developed cognitive models. Moreover, their findings suggested a shared representational hierarchy between CNN layers and visual cortex regions, where intermediate and late CNN layers correspond to intermediate and late visual processing regions, respectively. However, recent research that directly explored the emergence of brain-like hierarchy in neural networks trained to directly predict brain activity found evidence against the necessity of entailment hierarchy (St-Yves et al., 2023). Based on their results, the authors posit the *shallow brain hypothesis*, arguing that low-level representations may not be necessary as preprocessing stages for higher-level representations. Our work expands on this *shallow vs deep-brain* debate by employing Neural Architecture Search (NAS) to explore the emergence of early visual cortex-like representation in network architectures optimized to align with late ventral representation.

Previous studies have demonstrated that it is possible to identify CNNs with state-of-the-art classification performance by directly optimizing model architectures using methods such as reinforcement learning or genetic algorithms. For instance, (Xie & Yuille, 2017; Liu et al., 2018) leveraged genetic algorithms to 'evolve' architectures that outperform manually designed CNNs on MNIST and CIFAR-10 datasets. More recently, Mundt et al. (2021) employed NAS to identify network architectures that, even without gradient descent training, compute representations that enable classification performance comparable to fully trained deep networks by simply training a linear probe to predict the image label. Building on this, we optimized CNN architectures to predict cognitive representations across different regions of the ventral stream. We formulate a simple hypothesis in favor of the

[*]These authors contributed equally to this work
[†]Corresponding Author: sadiya@rz.uni-frankfurt.de

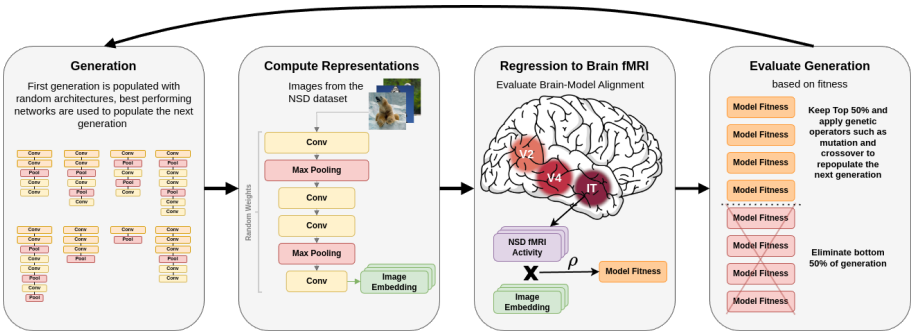

Figure 1: The evolutionary neural architecture search framework: Starting with a generation of neural architectures, for each model embeddings of all images in the shared NSD dataset are extracted. A ridge regression is then trained to predict the recorded fMRI activity and the correlation coefficient between the predicted and ground truth fMRI is calculated. The models are evaluated based on the mean correlation for all subjects, and the bottom 50% are eliminated. Finally genetic operations are used to repopulate the models for the next generation.

deep-brain model: Optimizing CNNs to predict late visual representations in the inferior temporal (IT) cortex would spontaneously optimize their alignment with representations found in the early (V2) and intermediate (V4) visual cortex regions in lower layers. Our results demonstrate that NAS can identify CNN architectures with better brain alignment than manually designed image classification models such as AlexNet, VGG16 and CORNet. Furthermore, we found that the optimal CNNs for predicting V2 and V4 representations were sub-networks of those optimized for predicting IT representations, indicating that these earlier representations might be necessary when trying to predict high level visual representations, therefore providing evidence in favor of the *deep-brain* model of the visual cortex. Finally, we also observed that architectures optimized to predict brain representations achieved competitive results when trained for image classification.

2 METHOD

We follow the standard evolutionary NAS methodology of allowing the genetic algorithm to perform selection, mutation, and crossover of the best networks in each generation to find an optimal architecture for brain alignment (Liu et al., 2018). In the following section, we present the specifics of this evolutionary architecture search. That is, we discuss the search space, the evaluation strategy, and the genetic operators used to evolve each new generation of CNNs.

2.1 SEARCH SPACE

Following Xie & Yuille (2017), we construct an initial generation of random individual networks, where each is a standard hierarchical stack of multiple convolution and max-pooling layers. We empirically found that adding linear layers to the search space did not improve brain-alignment and therefore excluded them to allow for faster convergence. We also limit the searchable hyperparameter ranges of the convolution layers kernel size (3 to 11), stride (1 to 4) and number of filters (64 to 512). Furthermore, the max-pooling layers kernel size range was 2 to 3. Finally, to further restrict the search space we also enforce the CNNs to always have a monotonically increasing number of filters across the layers, as receptive fields are expected to expand to integrate more conceptual information in deeper layers.

2.1.1 FITNESS EVALUATION

Adapting the approach of Mundt et al. (2021) we evaluate multiple randomly initialized versions of the same network, guaranteeing that the models are picked based on architecture, rather than a lucky weight initialization (the lottery ticket hypothesis). We use the Net2Brain toolbox to evaluate network performance Bersch et al. (2025): First, we extract the last layer image encoding and train a ridge regression to predict fMRI responses to the same image. We then calculate the Pearson correlation coefficient of predicted and ground-truth neural responses on a held-out test set. The co-

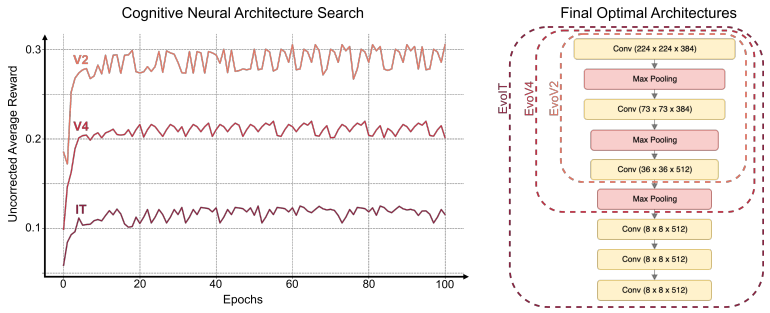

Figure 2: Left: The average rewards for the architectures optimized to predict V2, V4, and IT brain activations in each generation. Right: The optimal *EvoIT* architecture.

efficient is averaged across ten consecutive random seeds for each subject to generate model fitness. Finally, to speed convergence, in the first two generations we artificially increase the population size by evaluating every layer of each network (not just the last) and use the best performing sub-network in following generations. Following common practice, the random network weights were initialized using a Kaiming uniform distribution with a bias of zero.

2.2 Genetic Operators

After all the models in a generation are evaluated, we remove the bottom 50% of the CNN population based on fitness. We then create new offspring networks to repopulate the population. This is achieved using standard mutation and crossover genetic operators (Xie & Yuille, 2017).

Our mutation strategy employs three operators on each selected parent network: addition, modification, and removal. The addition operation introduces new layers while maintaining architectural validity, with special attention to channel dimensionality progression and layer-type constraints. Modification alternate between layer-type transformations (with probability $P = 0.3$) and parameter refinements ($P = 0.7$) in which architecture elements such as the kernel size are adjusted. The removal operation preserves network integrity by selectively eliminating layers while maintaining essential architectural elements (minimum depth of one layer and output size larger than one).

Our crossover operator implements a single-point crossover strategy in which architectures exchange structural information at a randomly selected position. The operation creates offspring by preserving the parent's layers up to the crossover point and inheriting the remaining layers from the second parent, maintaining architectural validity through constraint checking.

2.3 Dataset

The regression was trained using data from the NSD Dataset Allen et al. (2022), a large-scale fMRI dataset of 8 subjects viewing thousands of natural scenes. Specifically, we used a set of 872 images shared across all subjects. Following (Guo et al., 2025), we used a subset of five subjects (subjects 1,2,4,5,7) with a high signal-to-noise ratio (SNR). Moreover, we focused on the brain activity recorded in V2, V4, and the Inferior Temporal cortex (IT) as stand-ins for representations in the early, intermediate, and late ventral stream respectively. Specifically, IT activations were constructed by concatenating activations from the FFA, FBA, EBA and PPA regions Kanwisher & Dilks (2013).

2.4 Baseline Models

We compare the brain-alignment of our models against well known CNNs such as AlexNet and VGG16 which are often used as cognitive models (see Yamins et al. (2014); Guo et al. (2025)). Moreover, we also use CORNetS which was specifically designed to process information in a more brain-like manner Kubilius et al. (2019).

3 EXPERIMENTS AND RESULTS

The brain alignment results reported in this section are in terms of percent of variance explained relative to the lower noise ceiling. For representational similarity analysis we followed the standard lower noise ceiling calculation Nili et al. (2014). For the linear encoding based alignment score we used the regression method presented by Lage-Castellanos et al. (2019). We report performance for model layer with the highest alignment score.

3.1 EVOLUTIONARY SEARCH FOR COGNITIVE MODELS

We ran our evolutionary search three times for each brain region. Each run consisted of 100 generations with a mutation rate of 0.25 and a crossover rate of 0.5. The optimal architectures discovered were identical in all runs optimizing for the alignment with the same brain region save for minor differences in stride sizes, indicating that the identified networks are robust to noise in the evolutionary search process. Furthermore, the three optimal architectures for V2, V4, and the IT - which we coin EvoV2, EvoV4, and EvoIT - had high alignment with their respective brain regions, outperforming most baseline models (Table 1). The number of layers in the EvoV2, EvoV4, and EvoIT architectures was five, six, and nine respectively (Figure 2 right).

		V2		V4		IT	
		Reg	RSA	Reg	RSA	Reg	RSA
AlexNet	Random	8	5.15	5.14	3.09	1.29	0.40
	Trained	4.6	14.51	2.59	7.54	1.12	1.16
VGG16	Random	1.09	3.03	5.4	1.51	0.95	0.12
	Trained	**13.76**	16.12	4.82	6.91	0.75	0.85
CORNet	Random	2.38	1.52	2.33	0.91	0.65	0.33
	Trained	2.276	**17.82**	1.4	**9.99**	1.13	1.09
EvoV2	Random	11.9	5.99	6.15	3.47	0.95	0.41
	Trained	3.76	6.33	1.81	3.52	0.69	0.56
EvoV4	Random	11.98	3.94	6.5	2.47	1.2	0.41
	Trained	3.544	6.32	1.9	4.00	0.55	0.60
EvoIT	Random	10.64	3.90	**6.63**	2.70	**1.92**	0.43
	Trained	8.82	16.33	4.06	8.27	1.04	**1.18**

Table 1: Brain-model similarity as measured by linear encoding (Reg) and representational similarity analysis (RSA) across brain regions and model architectures with trained and random weights. Similarity is given calculated by normalizing the correlation coefficient using the lower noise ceiling. All Trained models use weights optimized on the CIFAR-10 classification (Section 3.3).

3.2 REPRESENTATIONAL HIERARCHY IN EVOLUTIONARY COGNITIVE MODELS

We formulated the following hypothesis in favor of the existence of hierarchical entailment across brain region representations: optimizing CNNs to predict activations in the IT representations will also spontaneously optimize the networks to learn V2 and V4 representations. To accomplish this we tested the brain-alignment of each brain region with each architecture (Table 1). Indeed, we find that EvoIT contained sub-networks that are competitive predictors of V2 and V4 (in fact, a subnetwork of EvoIT was the best in class V4 model). Specifically, the EvoIT layer that was most correlated with V4 was the third pooling layer (Figure 2 left). Moreover, manually inspecting the best performing architectures we observed that the EvoV4 and EvoIT architectures contained subnetworks virtually identical to EvoV2 and EvoV4 respectively. Overall, these results indicate that to compute representations similar to those found in the IT it is indeed beneficial to first compute representations similar to those found in V2 and V4.

3.3 Training Evolutionary Cognitive Models for Image Classification

CNNs designed to perform image classification are state-of-the-art cognitive models Yamins et al. (2014). Here we investigate if CNNs with architectures that were specifically optimized for brain-alignment can be trained to perform image classification. To achieve this we used the best architectures identified by the evolutionary search by adding a linear layer with Softmax on top of the CNN backbone and training the models with cross-entropy on the CIFAR-10 dataset for 7 epochs. To establish multiple baselines we also randomized the weights of the baseline CNNs and trained them using a similar procedure.

We observe that the classification performance increased for models optimized to align with later ventral regions, with EvoV2 having the lowest score, followed by EvoV4 and EvoIT. Overall, the performance of EvoIT was close to the performance of several baseline models (Table 2).

	EvoV2	EvoV4	EvoIT	VGG16	AlexNet	CORNet
Top-1 Accuracy	64.1	67.7	77.2	78.5	79.5	80.85

Table 2: Classification performance after training randomly initialized models on CIFAR-10

4 Discussion and Future Work

In this paper we used genetic algorithm based neural architecture search to optimize the architecture of the convolutional neural networks directly to be more brain-like. Through this framework we identified network architectures that had high similarity to various brain regions despite the lack of any gradient descent based training. Moreover, we found that the model optimized for similarity with late ventral stream areas contained subnetworks that were virtually identical to those identified as the optimal models for similarity with early and intermediate visual cortex representations. This finding directly contributes to the recent discussion regarding the hierarchy - or lack thereof - found across visual cortex representation (Yamins et al., 2014; St-Yves et al., 2023).

More broadly, the framework presented here is a potential new useful tool for computational cognitive neuroscience research. Previous research tackling questions regarding the architecture of the brain often followed a specific recipe: CNN architectures are modified in a controlled manner while the training data and function are held constant. Model-brain alignment is then measured to determine if the modification improves brain similarity, which would be taken as evidence that the modification constitutes an abstraction of a mechanism found in the brain (for example see Guo et al. (2025)). In contrast, the cognitive NAS framework presented here optimizes the model architecture directly, instead of relaying on handcrafted CNNs, which might introduce unwanted bias.

The initial results presented here highlight the potential utility of Cognitive NAS for the research community. However, the results of our experiments also raise multiple questions. Specifically, while the evolved networks were powerful encoding models, their performance was subpar when measured through representational similarity analysis. This might be due to our choice of using ridge regression for fitness evaluation. Interestingly, training the networks using the CIFAR-10 classification task improved brain-alignment as measured by RSA while lowering the regression score. This might indicate that the representation space of the evolved models lacks some structural elements that can only be learned through gradient descent training. Future work should carefully investigate the impact of the brain-similarity measure used during the NAS on the trajectory of the architecture search and the final networks identified as optimal cognitive models.

Acknowledgements

This work was partly funded by the German Research Foundation - DFG Research Unit FOR 5368.

REFERENCES

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25:116–126, 2022.

Domenic Bersch, Martina Vilas, Sari Sadiya, Timothy Schauml, Kshitij Dwivedi, Radoslaw Cichy, and Gemma Roig. Net2brain: A toolbox to compare artificial vision models with human brain responses. *Frontiers in Neuroinformatics*, 2025.

Manshan Guo, Michael Samjatin, Bhavin Choksi, Sari Sadiya, Radoslaw Cichy, and Gemma Roig. Predictive coding dynamics enhance model-brain similarity. *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2025.

Nancy Kanwisher and Daniel D Dilks. The functional organization of the ventral visual pathway in humans. *The new visual neurosciences*, pp. 733–748, 2013.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L K Yamins, and James J DiCarlo. Brain-like object recognition with high-performing shallow recurrent anns. In *Advances in Neural Information Processing Systems (NeurIPs)*, 2019.

Agustin Lage-Castellanos, Giancarlo Valente, Elia Formisano, and Federico De Martino. Methods for computing the maximum performance of computational models of fMRI responses. *PLoS Comput. Biol.*, 15(3):e1006397, March 2019.

Hanxiao Liu, Karen Simonyan, Oriol Vinyals, Chrisantha Fernando, and Koray Kavukcuoglu. Hierarchical representations for efficient architecture search. In *International Conference on Learning Representations*, 2018.

Martin Mundt, Iuliia Pliushch, and Visvanathan Ramesh. Neural architecture search of deep priors: Towards continual learning without catastrophic interference. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3518–3527, 2021.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Comput. Biol.*, 10(4): e1003553, April 2014.

Ghislain St-Yves, Emily J. Allen, Yue Wu, Kendrick N. Kay, and Thomas Naselaris. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature Communications*, 14:3329, 2023.

Lingxi Xie and Alan Yuille. Genetic cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1388–1397, 2017. doi: 10.1109/ICCV.2017.154.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.