Analysis and Discussion of Neural Fitting Results

Example Image from the Natural Scenes Dataset

Figure 1: Example NSD Image

This image from the Natural Scenes Dataset (NSD) shows a natural scene used in fMRI experiments. The NSD contains high-resolution natural images that subjects viewed while their brain activity was recorded. These images serve as the input for our neural fitting procedure.

Heatmaps of Neural Fitting Results

Heatmap showing normalized R^2 scores for the randomly initialized AlexNet model across different layers (rows) and brain regions (columns). The color intensity represents the proportion of explainable variance captured by each layer's representations.

Heatmap showing normalized R^2 scores for the ImageNet-trained AlexNet model. Note how the predictivity pattern differs from the random model, particularly for higher-level visual areas.

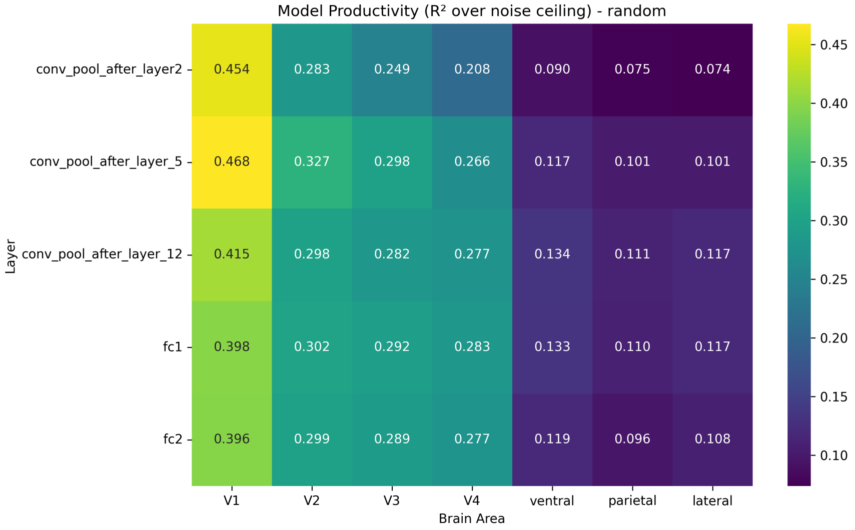

Figure 2: Heatmap for Random Model

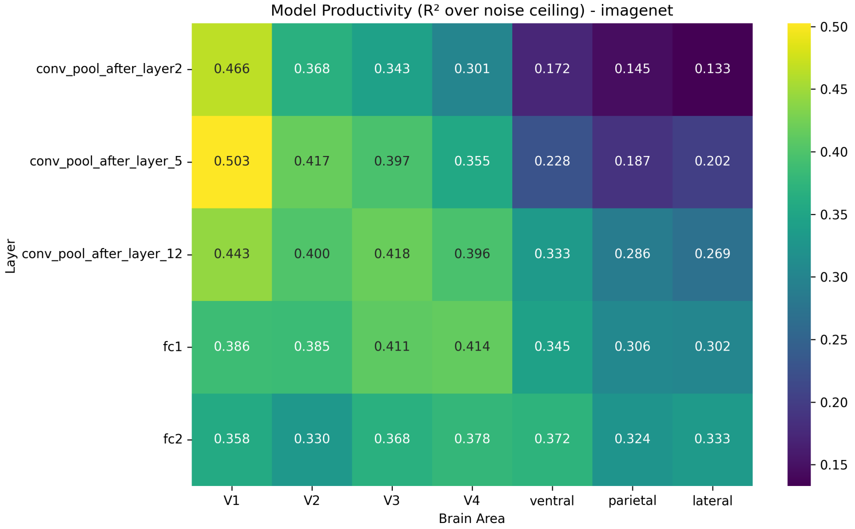

Figure 3: Heatmap for ImageNet-Trained Model

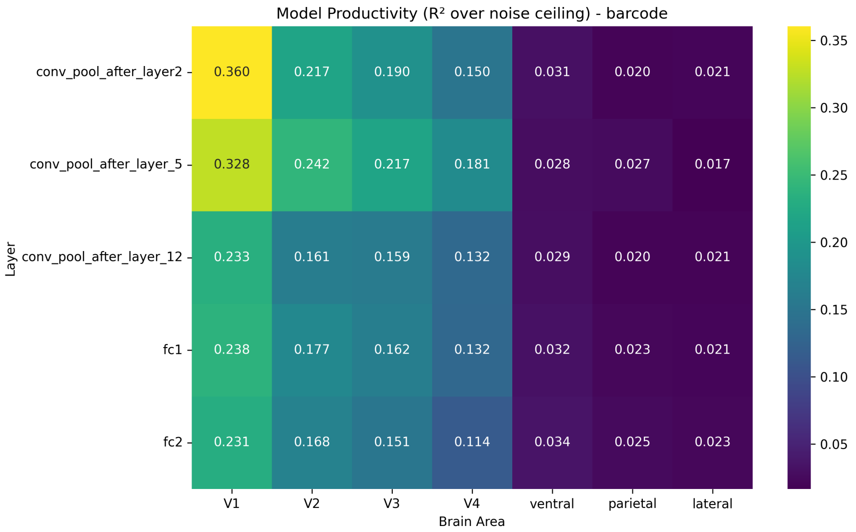

Figure 4: Heatmap for Barcode-Trained Model

Heatmap showing normalized R^2 scores for the Barcode-trained AlexNet model. This model was trained on a synthetic dataset of binary patterns rather than natural images.

1. Predictivity Patterns Across Layers

Based on the neural fitting results shown in the heatmaps above, I've compiled a comprehensive analysis of how different models predict neural activity across brain regions and network layers.

ImageNet-Trained Model

The ImageNet-trained model consistently demonstrates the highest predictivity across nearly all layers and brain regions. This suggests that representations learned for natural image classification align well with neural representations in the human visual system.

- **Early Layers (conv_pool_after_layer2, conv_pool_after_layer_5)**: Show strong predictivity for early visual areas (V1-V4), with normalized R^2 scores reaching ~0.45 for V1. This matches neuroscience literature showing that early CNN layers correspond well to early visual processing.
- **Later Convolutional Layer (conv_pool_after_layer_12)**: Demonstrates increased predictivity for higher-order areas like ventral, parietal, and lateral regions, while maintaining good prediction of early visual areas. This layer achieves normalized R^2 scores of ~0.41-0.42 for V3 and V4,

showing strong alignment with intermediate visual processing.

- **Fully Connected Layers (fc1, fc2)**: Show a slight decrease in predictivity for early visual areas but maintain strong performance for higher-order regions, suggesting that abstract representations learned for object recognition align with higher-level visual processing in the brain.

Random (Untrained) Model

The untrained model shows surprisingly good predictivity for early visual areas, especially in early layers:

- **Early Layers**: Achieve normalized R^2 scores of ~0.45-0.47 for V1, comparable to the ImageNet model. This suggests that the architectural biases of CNNs (local connectivity, pooling operations) inherently capture some aspects of early visual processing.
- **Performance Trend**: Predictivity decreases for higher-order areas and higher layers, indicating that while architectural biases capture low-level visual features, training is necessary to learn representations that align with higher-level brain functions.

Barcode-Trained Model

The Barcode model shows the lowest predictivity across most regions and layers:

- **Overall Performance**: Substantially lower than both random and ImageNet models, with normalized R^2 scores rarely exceeding 0.23, even for early visual areas.
- **Best Performance**: Occurs in early layers for V1 prediction (~0.23-0.24), but drastically underperforms compared to other models.

2. Barcode vs. ImageNet-Trained Model

The stark contrast between Barcode and ImageNet models reveals important insights about task-specific representation learning:

Task Nature Differences

- **ImageNet Task**: Classification of natural images requires learning hierarchical features that capture edges, textures, parts, and whole objects—features that align well with the ventral visual stream's processing.
- **Barcode Task**: Recognition of synthetic binary patterns with regular structure requires learning simpler features focused on detecting vertical edges and binary patterns. These features don't generalize well to natural scene processing.

Representation Transfer

- The Barcode model's low predictivity across all brain areas suggests that features optimized for artificial binary pattern recognition don't transfer well to explaining neural responses to natural scenes.

- In contrast, the ImageNet model's high predictivity indicates that features learned for natural image classification transfer effectively to predicting neural responses, suggesting shared computational principles between CNNs trained on natural images and the human visual system.

Layer-Specific Differences

- Early layers show the smallest gap between models, suggesting that basic visual features like edges are somewhat shared across tasks.
- The gap widens dramatically in higher layers, with the ImageNet model maintaining high predictivity while the Barcode model's performance collapses.

3. Performance of the Untrained (Random) Model

The random model's performance provides fascinating insights into architectural biases in neural networks:

Early Visual Areas Predictivity

- The random model achieves surprisingly high predictivity for early visual areas (V1, V2), with scores comparable to the ImageNet model in the earliest layers.
- This supports the idea that the basic convolutional architecture with random weights inherently extracts edge-like features similar to those processed in early visual cortex.

Architectural Bias

- The CNN architecture itself—with local receptive fields, weight sharing, and pooling operations—creates representations that naturally align with early visual processing, even without training.
- This architectural bias explains why random CNNs can predict early visual responses but fail to capture higher-level visual processing that requires learning.

Implications

- The high performance of random networks in predicting early visual responses suggests that the architecture of CNNs is inherently biased toward representations similar to those in early visual cortex.

- This finding aligns with research showing that random convolutional networks can extract meaningful visual features and perform above chance on simple visual tasks.

Conclusion

This neural fitting analysis reveals several key insights:

1. **Task-Optimized Features Matter**: The ImageNet model's superior performance demonstrates that learning features from natural images produces representations that better align with brain activity during natural scene viewing.
2. **Architecture Provides a Foundation**: The surprisingly good performance of the random model for early visual areas highlights how the CNN architecture itself captures some aspects of early visual processing.
3. **Task-Relevance Affects Transfer**: The poor performance of the Barcode model illustrates that learning features for artificial tasks that don't resemble natural images leads to representations that poorly predict brain activity.
4. **Representational Hierarchy**: The pattern of predictivity across layers and brain regions supports the hypothesis that CNNs and the visual system share a similar hierarchical organization, with early layers/regions processing low-level features and later layers/regions processing more abstract information.

These findings contribute to our understanding of both artificial neural networks and biological visual systems, suggesting that CNNs trained on natural images may serve as useful models of visual processing in the brain, while highlighting the importance of both architecture and task-specific optimization in developing models that capture neural computations.