Branch speacialziation:

Task representations in neural networks trained to perform many cognitive tasks

Analysis of Branch Specialization and its Application in Image Decomposition

Branch Specialization

Interpretability of artificial neural network models in artificial intelligence versus neuroscience

Feature attribution variance:

Improving performance of deep learning models with axiomatic attribution priors and expected gradients

This paper, featured in Nature Machine Intelligence, discusses how feature attribution methods for deep networks can be incorporated into training. This approach optimizes for a model whose attributions are aligned with certain priors or expectations. It's an exploration of how understanding and manipulating feature attributions can directly impact model training and performance (Nature Machine Intelligence).

LEARNING EXPLAINABLE MODELS USING ATTRIBUTION PRIORS

Basics:

Inception

Deeplift

ig