

Nordic AI Meet 2026

O2. Machine Learning & Intelligent Systems (Wed 10:15-11:15)

1. Deciphering neuroscientific knowledge encoded in brain imaging models using concept-based explainable AI

Authors: Esten Høyland Leonardsen^{1,2}; Andreas Dahl^{1,2}; Øystein Sørensen¹; Lars T. Westlye^{1,2}

Affiliations: 1. Department of Psychology, University of Oslo, Oslo, Norway; 2. Centre for Precision Psychiatry, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway

Abstract: Artificial neural networks (ANNs) can accurately differentiate patients with various brain disorders based on Magnetic Resonance Imaging (MRI), emphasising their potential as tools to further our understanding of these enigmatic conditions. However, their opaque nature makes it challenging to translate the patterns they find into scientific and clinically actionable insights. To overcome this, explainable artificial intelligence (XAI) renders the decision processes of ANNs interpretable to humans. In clinical neuroimaging, most XAI applications have relied on feature attribution to uncover what brain regions underlie the predictions of a model, without revealing what it finds there. By explaining the computational processes of ANNs in human-interpretable terms, semantic XAI offers a complementary path to revealing the neuroscientific insights these models encode.

Here, we developed a model dissection pipeline to describe the inner workings of a dementia classifier with a human-interpretable vocabulary. Using a large language model, we compiled a curated set of linguistic concepts representing dementia-related and incidental radiological findings based on two medical ontologies and a brain atlas. We then used Prima, a contrastive language-image pretrained (CLIP) model trained to align brain MRIs with neuroradiological reports, to annotate MRIs from 11,462 participants (6,890 healthy controls, 4,572 cognitively impaired) with corresponding concept labels. Next, we finetuned SFCN-multi, an open-source pretrained convolutional neural network, to predict all-cause dementia. By calculating concept activation vectors (CAVs), we assessed whether known radiological markers were encoded in the layers of the model, and applied TCAV to quantify their influence on its predictions. By building a post-hoc concept bottleneck model (CBM) to differentiate frontotemporal and Alzheimer's dementia, we examined whether the model had discovered emergent, clinically relevant structure that could be dissected along interpretable dimensions.

In preliminary results, our dementia classifier accurately distinguished patients from controls (AUC>0.9), and both CLIP encoders produced seemingly coherent embedding spaces. However, closer inspection revealed that our concept embeddings showed patterns consistent with representational collapse, which propagated into both cross-modal alignment and downstream analyses. As a result, both the CAVs and the CBM latched on to a low-dimensional confounding signal rather than concept-specific content. Overall, our concept embeddings contained coarse dementia-relevant signal, but lacked the resolution required to perform model dissection. Applying corrections for such structural distortions in high-dimensional representational spaces appears to recover concept-specific signal, and emerges as a necessary step to enable knowledge discovery in domains like neuroimaging, where feature redundancy causes confounds to dominate the data manifold.