Concept Activation Vectors (CAV)

💡 High-level Concepts

Testing for abstract medical concepts in neural networks

➡️ Direction Vectors

Finding concept directions in model activation space

🧪 Sensitivity Testing

Measuring model response to specific medical concepts

🔬 Medical Concepts

Detecting learned patterns like 'inflammation' or 'tumor characteristics'