doi:10.1631/FITEE.1700808

题目：深度学习中的视觉可解释性

概要：总结了近年来在理解神经网络内部特征表达和训练一个具有中层表达可解释性的深度神经网络上的相关研究工作。虽然深度神经网络在众多人工智能任务中已有杰出表现，但神经网络中层表达的可解释性依然是该领域发展的重大瓶颈。目前，深度神经网络以低解释性的黑箱表达为代价，获取了强大的分类能力。我们认为提高神经网络中层特征表达的可解释性，可以帮助人们打破众多深度学习的发展瓶颈，比如，小数据训练，语义层面上的人机交互式训练，以及基于内在特征语义定向精准修复网络中层特征表达缺陷等难题。本文着眼于卷积神经网络，调研了：(1)网络表达可视化方法；(2)网络表达的诊断方法；(3)自动解构解释卷积神经网络的方法；(4)学习中层特征表达可解释的神经网络的方法；(5)基于网络可解释性的中层对端的深度学习算法。最后，讨论了可解释性人工智能未来可能的发展趋势。

关键词：人工智能；深度学习；可解释性模型