可视化解释深度学习

—— 计算机视觉

倪子烜 (11821025)

摘要

本文简单综述了近年来在理解神经网络表示和学习神经网络方面的研究。尽管深度神经网络在各种任务中表现出了卓越的性能, 但可解释性始终是深度神经网络的致命弱点。目前，深度神经网络通过低解释性的代价获得了高分辨率的结果，然而随着未来的发展，高解释性将帮助人们打破现今深度学习的许多瓶颈（如弱监督学习，人机交互学习以及网络修复）。

现在的可解释性主要集中在卷积神经网络 CNNs（convolutional neural networks）[5] 方面，本文主要回顾了可视化 CNN 的表示，分析预训练 CNN 的方法，诊断预训练 CNN 的方法以及模型的中到端学习方法。最后，主要讨论了未来我认为的人工智能趋势。

1. 介绍

卷积神经网络 CNN(Convolutional neural networks) [5] 在物体识别检测等众多任务中达到了很高的精度和准度。然而端到端的学习策略使得 CNN 成为了一个无法解释的黑箱, 除了模型最终的输出结果以外,CNN 内部的网络结构和逻辑都是我们现今无法理解和掌握的。最近几年, 大量的研究工作者已经意识到了一个高度的模型可解释性是非常重要的对于模型的理论与实践的发展，同时也发展出了许多可解释的知识表达模型。

现如今, 我们可以粗糙的定义可解释方法分为下面 5 个研究方向。

- 使用可解释的网络层去可视化 CNN。该方法主要融合在训练好的 CNN 中最大分数的单元图片或者将输入图片的卷积层特征图转换回原输入图片。
- 分析 CNN 特征。相关工作主要是诊断不同类别的 CNN 特征空间或者是发现潜在的卷积层特征缺陷。
- 解开 CNN 中的混合特征编码。该方向的研究主要是尝试分离卷积层中的复杂特征并转换网络表达为可解释图。
- 搭建可解释的模型。例如可解释的 CNN 模型 [21], 胶囊网络 [12], 可解释的 RCNN 以及 InfoGAN [3]。
- 通过人机交互进行语义水平的中端学习。一个清晰语义解释 CNN 的表示可能会进一步使弱监督神经网络的“中端”学习成为可能。

上面的 5 个研究方向中, 可视化 CNN 的表达是最直接的方法去解释网络的意义。网络的可视化也提供了分析 CNN 特征的许多有效方法。拆解预训练的 CNN 特征并学习可解释的特征表示为最佳算法提出了更大的挑战。同时, 可解释与分析网络的表示也是弱监督学习和中到端学习的开始。

模型可解释的价值在于高层卷积网络由于其复杂性很难被理解，通过语义表示可以帮助人们提高对模型预测结果的信任度。举个例子，在物体检测过程中 CNN 也许可以分辨一张图片中人是否涂口红，但是判断的依据也许并不是口红本身，而是图片中人的性别和是否化妆。这样的判断依据很容易导致现实数据中的误判 [22]。因此，人们时常无法相信 CNN 的预测结果除非 CNN 可以语义话他的判断过程或者可视化他的预测逻辑（例如图片中哪一个部分被用于判断）。

除此以外，将可解释性用于中到端学习或者查漏神经网络都将帮助人们明显降低人工标注的数量。甚者，依据神经网络的语义表达，将多种 CNN 融入语义层的普适网络用于不同任务下的知识表达。

2. 可视化 CNN 表达

可视化 CNN 中的滤波器是最直接的方法去探索神经单元中隐藏的特征。现已出现了多种可视化网络方法。

最初是依据梯度的方法例如“Visualizing and understanding convolutional networks” [19]；“Understanding deep image representations by inverting them” [8]；“Deep inside convolutional networks: visualising image classification models and saliency maps.” [13]；其中“Striving for simplicity: the all convolutional net.” [14] 是主流的网络可视化方法。这些方法主要是通过计算 CNN 单元关于输入图片的梯度分数来可视化结果，通过梯度最大化神经单元的分数从而评估所得图片的表达是否合理。在这期间 Olah 等人 [10] 提出了 *distill* 技术编码预训练 CNN 下的不同卷积层的可视化图片。

随着技术的发展，Dosovitskiy 等人提出了另一种典型的可视化 CNN 表示技术既上卷积网络 [4]。上卷积网络是将 CNN 的特征图转化为图片但是他并不能直接的显示图片和特征图之间的联系。与梯度方法相比，上卷积网络并不能确保可视化后的结果反应了 CNN 内部的实际表达。与其相似的还有 Nguyen 所提出的关于融合图片语义的先验方法 [9] 用于对抗生成网络。

除此以外，[23] 提出了一种精确计算特征图中接收到的神经激活领域。论文表示，实际接收到的神经领域比使用滤波器尺寸计算出来的理论接收领域小很多。接收邻域的精度帮助科研者理解滤波器的表达。

3. 分析 CNN 的表示

一些方法不但可视化了 CNN，同时还分析了 CNN 的表示从而尝试理解 CNN 的特征编码。该研究方向可以大致分为 5 个小方向。

第一个便是尝试从整体分析 CNN 的特征。Szegedy 等人探索了每一个滤波器的语义 [15]。Yosinski 等人分析了卷积层内部的滤波器表达转换 [18]。[7] [1] 则计算了预训练 CNN 下的不同类别的特征分布。

第二个则是抽取对网络输出直接贡献最大的图片区域从而解释网络关于该标签的表达。这和 CNN 的可视化很相似. 其中最具代表性的就是 LIME 模型 [11]，该方法尝试抽取对于网络输出最敏感的图片区域。同时，Zintgraf,Kindermans,Kumar 等人则尝试可视化输入图片中对 CNN 贡献最大的区域。Wang 和 Goyal 等团队则努力解释 VQA 中神经网络内部的逻辑。

除了解释以外，尝试评估 CNN 特征空间的缺陷也是分析神经网络特征的一个热门方向。Su,Koh,Szegedy 等人则尝试去计算 CNN 的对抗样本，例如评估尝试改变 CNN 模型预测结果的最小噪声。Koh 等人提出的 CNN 影响函数则可以帮助计算对抗样本，其产生的训练样本可以反向“攻击”训练中的 CNN，固定训练集，更进一步的纠正 CNN 的表达。

第四个研究方向则是依据网络的特征空间分析精调网络的表示。例如给一个预训练的 CNN 物体检测模型，提出方法去发现 CNN 在弱监督模式下的未知面 [6]。

最后一个方向则是发现 CNN 中的潜在错误表示。如下图所示，展现了 CNN 在预测面部属性时的偏执表达。

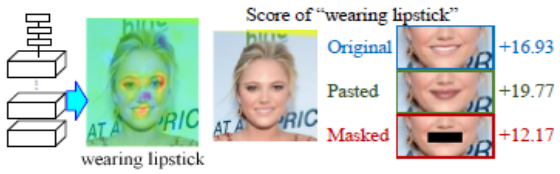

Figure 1: Biased representations in a CNN [Zhang *et al.*, 2018b]. Considering potential dataset bias, a high accuracy on testing images cannot always ensure that a CNN learns correct representations. The CNN may use unreliable co-appearing contexts to make predictions. For example, people may manually modify mouth appearances of two faces by masking mouth regions or pasting another mouth, but such modifications do not significantly change prediction scores for the *lipstick* attribute. This figure shows heat maps of inference patterns of the *lipstick* attribute, where patterns with red/blue colors are positive/negative with the attribute score. The CNN mistakenly considers unrelated patterns as contexts to infer the lipstick.

当一个属性在训练数据中经常和另一个特殊的可视化特征一起出现时，CNN 也许会依据这个特征来判断是否有那个属性。如上图所示，CNN 以女人是否化妆来判断其是否涂口红。

4. 通过可解释图和决策树来分析 CNN 表示

与先前的可视化与网络表达分析相比，将 CNN 的特征拆分为人们可解释的图表示为网络的解释提供了新的方法。Zhang 等人提出拆分预训练的 CNN 中的特征并使用图模型表示 CNN 内部隐藏层的语义层次 [20]。如下图所示：

Figure 2: Feature maps of a filter obtained using different input images [Zhang *et al.*, 2018a]. To visualize the feature map, the method propagates receptive fields of activated units in the feature map back to the image plane. In each sub-feature, the filter is activated by various part patterns in an image. This makes it difficult to understand the semantic meaning of a filter.

在 CNN 内部高层的卷积层时常代表混合的图像。比如某一层滤波器也许会同时被物体的头部和尾部激活。因此尝试将各个激活部分都来自哪些物体部分拆解出来有助于帮助我们了解有多少可视化图像被 CNN 滤波器记住以及滤波器上哪些被激活的部分描述了物体的这个部分，两个共同激活的部分有什么空间关系。

决策树的方法与可解释图理念一致，只是决策树的方式并不是用于分类，而是为了解释 CNN 分类的逻辑。当我们给一张输入图片的时候，我们使用 CNN 去预测。而决策树则告诉我们在卷积层中的哪个滤波器被用于这次预测，以及它对预测做出的贡献有多少。

5. 使用可解释的表示去训练神经网络

之前的方法主要在理解预训练的神经网络下的 CNN 特征。而这个方法则意在学习可解释的神经网络，使得神经网络内部结构可解释化。代表工作有 [21] [16] [17]。三种方式所使用的方法皆不相同。[21] 方法是将模型损失用于特征映射规范化，以表示特定的对象部件。而 [16] 则是提出了定性的依据 R-CNN 的可解释模型用于物体识别。Hinton 提出的 [17] 则尝试设计新的神经单元命名为胶囊。尝试通过新的模型来获得一举多得的效果。

6. 可解释性的评估标准

任何领域的开拓都需要一个标准。模型可解释性的评估标准也是判定模型可解释能力的重要标准。然而网络可解释性的判定并不像普通的任务可以通过比如物体位置，物体类别等宏观的标准来判定。直到现在，被讨论最多的两种可解释性的判断标准一是滤波器的可解释性，二是位置不确定性。

6.1. 滤波器的可解释性

最早是由 Bau 在论文 [2] 中提出了六种 CNN 滤波器的类型分别为“对象、部件、场景、纹理、材料和颜色”。评价指标衡量的是滤波器神经活动的图像分辨率接受域与图像上像素级语义标注之间的适应度。例如，如果一个滤波器的神经激活的接受域通常通过不同的图像与特定语义概念的真实图像区域高度重叠，那么我们可以认为滤波器代表了这个语义概念。

6.2. 位置不确定性

这一度量指标由论文 [20] 提出。其目的在于评估 CNN 过滤器与对象部件表示之间的适应度。如下图所示：

Figure 12: Notation for the computation of a filter's location instability [Zhang *et al.*, 2018a].

通过计算推断位置与实际位置之间的偏差来求得位置不确定性。

7. 总结和未来发展

可解释性作为深度学习必经的一环，也是极其重要的一环。可视化神经网络的单元是早期理解网络表达的开始点。之后人们逐渐发展各种方法去分析神经网络的特征空间并尝试诊断神经网络内部的错误。目前，将卷积层的混沌表示分解为图形模型和/或符号逻辑已成为打开神经网络黑箱的一个新兴研究方向。提出了一种将训练好的 CNN 转换为解释性图的方法，该方法在知识转移和弱监督学习方面具有显著的效率。

端到端学习可解释的神经网络，其中间层编码可理解的模式，也是一个未来的趋势。可解释的神经网络已经被开发出来，其中高卷积层中的每个过滤器代表一个特定的对象部分。

然而现在的可解释性争对的 CNN 网络也许并不是最正确的网络模式，未来的可解释性将要处理的应该是修正网络模式，通过模型可解释性减少训练时间以及数据量，甚者帮助人们更好的理解投资判断与选择。

除此以外，现在的模型更多的是在大概率下融合所有训练数据对其进行概率统计。这与人类的判断方式完全不同，人类则是通过少量数据确定物体特征之后对物体举一反三从而不断学习和发现。为此在未来的可解释性发展过程中能否通过可解释的度量来调整模型的方向，实现中到端的训练模式也将是一个值得研究的方向。

在未来，我们相信中端学习将会是一个不断发展的基础研究方向。此外，基于可解释网络的语义层次结构，在语义层调试 CNN 表示将创建新的可视化应用程序。

参考文献

[1] Mathieu Aubry and Bryan Russell. Understanding deep features with computer-generated imagery. In IEEE International Conference on Computer Vision, 2015.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6541–6549, 2017.

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in neural information processing systems, pages 2172–2180, 2016.

[4] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4829–4837, 2016.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[6] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[7] Yao Lu. Unsupervised learning of neural network outputs. 2015.

[8] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5188–5196, 2015.

[9] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4467–4477, 2017.

[10] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. Distill, 2(11):e7, 2017.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144. ACM, 2016.

[12] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In Advances in neural information processing systems, pages 3856–3866, 2017.

[13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.

[14] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806, 2014.

[15] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. Computer Science, 2013.

[16] Tianfu Wu, Wei Sun, Xilai Li, Xi Song, and Bo Li. Towards interpretable r-cnn by unfolding latent structures. arXiv preprint arXiv:1711.05226, 2017.

[17] Edgar Xi, Selina Bing, and Yang Jin. Capsule network performance on complex data. arXiv preprint arXiv:1712.03480, 2017.

[18] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? Eprint Arxiv, 27:3320–3328, 2014.

[19] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In European conference on computer vision, pages 818–833. Springer, 2014.

[20] Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[21] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8827–8836, 2018.

[22] Quanshi Zhang, Wenguan Wang, and Song-Chun Zhu. Examining cnn representations with respect to dataset bias. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[23] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. arXiv preprint arXiv:1412.6856, 2014.