Could a neuroscientist understand a microprocessor?

http://freemind.pluskid.org/misc/could-a-neuroscientist-understand-a-microprocessor

posted on Free Mind on May 26, 2016
generated with pandoc on May 27, 2016
category: MISC

tags: Thoughts, Fun

最近看到一篇有趣的文章叫做 Could a neuroscientist understand a microprocessor? (DOI: 10.1101/055624)。讲的是用神经科学的分析方法来把计算机微处理器当做一个黑盒子来研究。这让我想起之前听过的一个 talk，里面有提到这篇叫做 A search for life on Earth from the Galileo spacecraft 的文章。

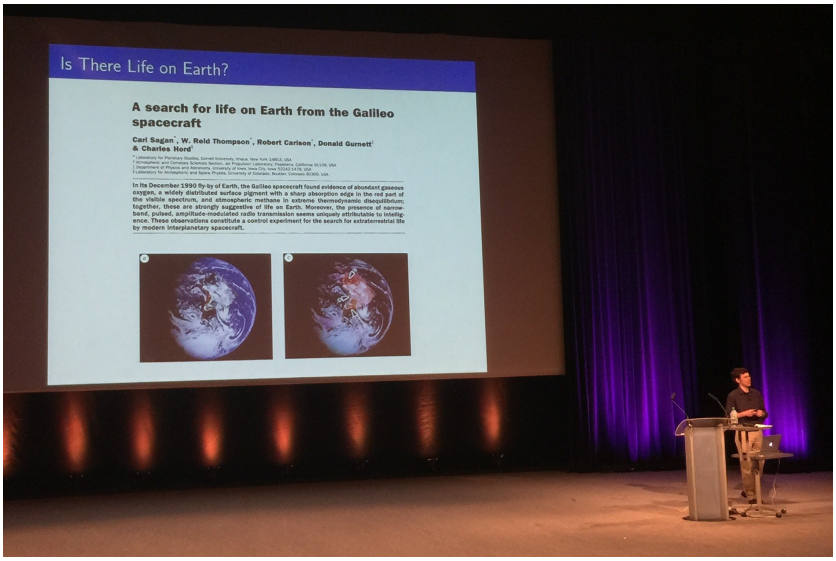

Figure 1: .....

就是说人们开发了许多技术和手段来探测外星生命的存在，然后终于有人想起来拿这些方法来对地球做了一下探测，看是否真的能探测到生命的存在，或者说，从另一方面来看，看这些技术和手段是不是真的有用。看似非常自然的基本测试，好像却还是比较容易被忽略的样子。

而这篇 neuroscience 的文章，其实也是类似的东西。首先 neuroscience 是一个比较新的学科，我自己的不完全理解它很大一部分内容是研究大脑的工作原理。所谓 science，差不多就是一个设计实验、观察结果、总结规律、建立理论、行为预测、结果检验的迭代过程。比如牛顿看到苹果从树上掉下来，于是建立了万有引力理论，用于刻画苹果的受力，同时还建立了牛顿运动定律之类的，用于刻画在受到给定的力的情况下苹果的运动。我们之所以接受这套理论是因为它能够进行 “行为预测”，比如它能够描述梨从树上掉下来，桃子从树上掉下来，或者我们可以设计 controlled experiment，让铁球从树上掉下来，观察运动轨迹，验证理论的正确性。当然相比于 “预测”，一套理论更重要的作用也许在于能帮助我们 “理解”，就是了解事情的来龙去脉，因果关系。

但是与此同时，Science 之于 Math 的一大特点就是基本上 “只能证伪

不能证实”。就是说如果我观察到和理论不符合的实验结果，就可以证明理论是不正确的，但是不论我做出多少符合预期的观测结果，都不能严格地证明理论是 “正确的”。比如后来爱因斯坦建立相对论，因为出现了一些牛顿理论无法很好地解释的现象。而爱因斯坦的理论是否又是 “正确” 的呢？或者说后来的其他理论。大概这是一个永无止境的过程。

而神经科学也面临着同样的问题，比如我们通过对脑部的 activity 进行测量，发现当我们看到人脸的时候，有一部分的神经元会有比较强烈的活动，于是我们可以建立一个关于脑部对于人脸专用的处理单元的模型或者理论，或者甚至关于 Grandmother cell 的理论。在比如通过对某些脑部由于意外受到部分损伤的病人的研究，发现某些部位损伤之后病人不能说话了，另一些部位损伤之后病人不能造句了，于是我们可以建立一个脑部各部分功能映射的模型等等。这里说得有点夸张和简化，但是大致差不多就是这样子。

从 “预测” 和 “理解” 的角度来看的话，神经科学的目的大概可以看成是 (1) 构建一个计算模型（或者实体计算机）能够实现大脑的功能；(2) 一套完整的理论用来描述大脑计算的来龙去脉。我自己的感觉是第一个目标大概是要容易很多的，也许不远的将来人们就能够通过海量的数据和计算训练出一个神经网络或者什么其他模型来，可以达到简单动物大脑类似的功能，然而这并不等价于同时实现了 (2)，因为这个训练出来的模型也许就是一个很复杂的难以 “分析” 的东西，到头来我们并不知道它到底是怎么回事。一个极端一点的玩笑就是，其实要实现 (1) 很 “容易”，只要 follow 一些固定程序，十月怀胎生一个小孩，你就算是 build 了一个 powerful 的大脑了，但是你还是不知道这个大脑是怎么 work 的。

但是所谓 “理解” 究竟是什么意思呢？感觉好像一个大脑是一件非常复杂的艺术品，似乎并不能像许多物理上的现象那样能用比较 clean 的几个公式就能完整刻画。并且如果只能证伪不能证实的话，似乎也很难讲什么是真正的 “理解”。

所以这里计算机微处理器就登场了。虽然还达不到人脑的复杂度，但是一个现代的 CPU，或者说一台完整的电脑，显然已经是非常复杂了。并且幸运的是它是我们人类自己建造的，我的意思是说，所有的模块、功能、层次抽象从头到尾都是人类设计出来的，而不是在一个全连接的电路板上通过 “gradient descent” 之类的方法 “train” 出来的。在这个情况里，我们知道了计算机这样一个复杂系统，虽然极其复杂，但是并不是无法 “描述” 或者 “理解” 的，虽然也许要用上很多本大部头在加上 EECS 专业的学生数年时间的学习才能得到一个相对不那么粗略的系统概貌。第二点就是我们有一个 ground truth 在，所以我们知道当我们得到一些结论或者理论的时候，到底是正确的还是错误的。当然还有很多其他优点，比如在模拟器上能够做非常精确又不影响原始功能的测量（反过来要在不破坏大脑功能的情况下做很精确的测量其实是比较难的），还有基本上可以免费生成无限量的实验数据等等。

所以说，回到文中最开始提到的这篇论文。作者的做了一些有趣的

实验，就是用神经科学里常用的一些实验设计和数据分析方法来分析一个游戏机板子。比如通过观测某些 transistor 会对某些像素有强烈的反应之类的来对 transistor 的功能进行刻画。再或者通过 “杀死” 一部分的 transistor 看是否会影响某些功能，来判断某些 transistor 或者 “区域” 的功能，比如作者发现禁用掉某一些 transistor 之后 “大金刚” 游戏就启动不起来了，那是否这些 transistor 就是大金刚专用 transistor 呢？当然我们知道微处理器本身是 general purpose 的结构，而逻辑存在于软件之中，可能就会觉得这样的结论不一定是正确的：比如也许它们其实是某些关键运算（比如加法）的 transistor，导致程序不能正确运行了，或者把磁盘之类的也考虑进去的话，也许刚好是存储大金刚程序的那部分 “neuron”，所以程序不能正确运行了。等等。

> *We find that many measures are surprisingly similar between the brain and the processor and also, that our results do not lead to a meaningful understanding of the processor.*

且不说文章里用的神经科学的实验设计和分析方法是否足够完整和严格，也不说它所得出的结论是否合理，就这个问题或者说设定本身来说其实是非常有意思的。虽然人脑和电脑的工作原理也许会有 fundamental 的不同，但是如果我们能有办法以黑盒子的方式从外部 reconstruct 出一整套关于电脑的工作原理的完整描述出来的话，那我们也就能很好地理解人脑了吧！