* 智能：如何生存

就我个人而言，入门任何知识的第一步就是知道目的。它能够做什么，我什么时候需要它。所以第一部分，我会先简单描述智能，为得是让大家清楚深层学习到底能够做什么。为什么画面识别，语音识别，自动驾驶，机器翻译，大量的应用都可以使用深层学习。同时也想消除媒体对大家造成的恐惧。

先想象一个生存游戏

规则是，四个格子分别代表环境的四种状态，靠近与否会产生两个后果，消失或存活。

个体全部消失则game over。

每个个体都有一个从环境到行为的关联f，遇到不同情况，会产生对应的行为，而这种f是随机生成的。只有在该随机生成的f恰好能够让他不靠近危险的环境时，该个体才能存活。

然而即便我们非常幸运的随机生成了正确的f，单个个体最终依然会消失。原因在于环境会变。于是大自然是利用大量的个体不断繁殖，并且该繁殖过程并非完美复制，而是会产生变异，变异的个体就有机会适应变化后的环境而存活下来，无法适应环境的关联会被筛选掉。随后继续大量繁殖。筛选后的结果便是进化。

游戏的关键在于，关联的更新速度要大于环境的变化速度，才不至于被全部筛选掉。所以自然界的单细胞生物大量繁殖的目的在于增大可供筛选的样本空间。

进化的对象不是个体，也不是基因，而是智能关联f，进化的过程是智能关联不断被筛选的过程。寻找合适的f的过程就是学习，自然选择是一种动态学习的方式。

用霍金的话说，智能是根据变化而产生相应变化的能力。

其中有三个要素，学习，寻找到合适的关联f，延续，f的存储，执行，依照f产生相应的行为。对低等生物而言，这三个要素分别靠自然选择，蛋白质识别，DNA来实现。

同时生命还有不断链接形成更高只能提从而增强智能能力的倾向，比如蛋白质，到细胞，到组织，到器官，到个体，到国家，更能会越来越强，生存下去的可能性也会适当增加。

环境的变化也是难以预测的，所以进化并没有方向，但是却有增加差异性的趋势。差异性的增加会让种群更容易适应环境的变化。于是大自然产生了有性繁殖。问题也相应产生，如果生命不可以移动，那么个体只能与周围的其他个体进行繁殖，有性繁殖的差异性能力被限制。为了洗牌，大范围移动成为了必然需求。可大范围移动同时也带来了另一个问题，环境会随着移动而变化，个体在上一环境中所学习到的关联，在下一个环境并不适用。这时该如何保障个体不触雷？

由于增加了移动，感官等能力，个体的生长周期变长，再也无法承担由自然选择去试错所带来的代价。于是对于大范围移动生物而言，他们需要原有智能的基础上增加了另一项能力：预测。 关联f不再是简单的从环境到行为的关联。而是可以从过去事件到未来事件的关联。生物会利用意识在大脑中模拟自己与环境的关系，通过预测下一步，甚至下几步会发生什么，从而进行决策。

神经网络可以建立任意两个空间的关联，当x代表的是画面，而y代表的是画面在大脑中的概念时，神经网络所完成的就是画面识别。当x代表所说的话，而y代表将要说的话，那么神经网络所完成的就是语言模型，当x代表英语，而y代表中文，神经网络所完成的就是机器翻译。神经网络的作用就是建立两个空间的关联。在以后的活动中就可以利用所建立的关联，通过输入得到我们想要的输出。

而一个人的一生就是在不断建立自己所生存的环境的各式各样的预测模型。茶杯坠落会发生什么，有恐高症的人站在高处，会预测自己坠落的后果。随着环境的复杂，所建立的预测模型需要跟着扩展和更新。比如投资项目，未来几个月的股票、房价、汇率的变化等。

但是光靠神经网络是无法支持一个可以大范围移动生命体存活的。智能除了建立关联外还有何时搜集数据，何时更新已建立的模型，何时使用哪个模型。这些都是需要在已有智能的基础上再次迭代形成更高级的功能，而这种功能多来源意识。

人脑的预测实际上是双向的，会时时刻刻预测即将发生的事情，并且与实际发生的事情进行比较，只有那些与自己预测不符合的事件才会得到意识的注意。比如你不会注意你每天都走的楼梯，但是当某个阶梯突然高出3厘米时，你却很容易察觉。这种与自己所建立的模型冲突的事件会更容易被记住并且作为训练数据搜集起来。

另外人的一生都在学习，因为世界不断变化。并不存在小孩子比成人学习能力强的一说。之所以造成这种现象的原因是人们以来自己的世界模型，孩子会什么都接受，没有辨别能力。而成人有保护机制，防止被人忽悠。但是也会更新已建模型的。比如说服一个成人，较好的做法是描述一个事件，这时他会不由自主的预测结果。但当你告诉他，他的预测是错误的时候，他就得到了“原来自己的已有模型并不可靠的信号”，并关闭抵制学习的保护机制来更新已有模型。智能始终都是并行迭代形成的，唯独意识不可以同一时间出现在两个地方，原因之一在于意识必须要在某个时刻决定应用哪个关联。

所以不用被媒体忽悠，认为目前的人工智能有取代人类的危险。真正有危险的并不是人工智能本身，而是如何使用人工智能的人类，如果将该技术用于开发武器，那样所产生的战争才是毁灭人类的原因。

我尽可能少的给大家列出关键点，并不意味着其他内容不重要。

深层学习是建立两个空间的单向关联，这里特指有监督学习。只要有正确的数据，深层学习几乎可以应用在所有行业上。很多时候我们需要根据特定的情况，产生合理的决策。

很多深层学习所接受的输入往往是观察，比如燕子地飞，天空多云，空气潮湿等observations。

其输出是造成这类现象的起因。

这些所有的输入输出，以及变化，都可以用矩阵来表示。所以线性代数是告诉你该如何表达这些内容的学科。

不断变化的世界使我们产生时间观念。正确描述事物状态及其不同时间下的变化至关重要。我们知道在三维空间下如何描述物体的位置。然而除了长宽高，世界上还有很多决定事物状态的因素。如决定股票价钱的因素、决定天气的因素。这些因素又该如何合理的描述？线性代数给了我们答案。

通过线性代数，我们知道了该如何描述事物状态及其变化。遗憾的是，对一个微小的生物而言，世界并非确定性的，由于信息量的限制，很多事物是无法确定其变化后会到达哪种状态。然而为了更好的生存，预测未来状态以决定下一刻的行为至关重要。而概率给我们的决策提供了依据。

寻找合理解释数据的关联f 是所有机器学习的目标，但是为何偏偏深层学习如此的高效？这一部分的理解很大程度上决定了你对神经网络的设计。

DNA是基础指的是，DNA是该星球上智能的基础，它既可以稳定存在，又可以和蛋白质外加自然选择完成职能的三个要素。但是其他的如果有生命体，可能就会有完全不一样的实现该三要素的方式，存储媒介也不一定是DNA。同样，我们想构造一个智能体时，不管我们利用何种手段，只要能够完成那三个要素即可。比如我们现在就是利用计算机来进行关联f的存储，利用机械来完成f的应用，利用类似神经元的方式来寻找关联f。人们会发现人工智能非常像人类智能，之所以相似是因为二者的目的相同，但是却不一定利用相同的手段。

学习是需要从历史经验中训练，希望可以在未见过的任务上具有良好表现。

就拿高考来说，我们通过做历年的真题来训练自己的解题方法，也就是关联f。

训练的方式是不断的做题，并且对照正确答案。等到高考时，当看到新题目x时，希望可以利用自己所训练的关联f，来得到正确答案y。我们真正关心的并不是已经做过的历年真题，而是高考。

高考有多难相比不用我说了。

另一方面的难点在于无法见到所有情况。

假如说，高考只有4道题判断题，只需要记住4到题的对错，高考就可以拿满分了。

可是实际的情况是，高考所能出的题是无限的。而我们可以训练的历年真题确实有限的。

又比如要识别猫的图片，猫会有各种各样的形态，表情，颜色，大小，各式各样的变体。该我们又无法穷举所有的猫，该如何从有限的猫图片样本中训练出较好的能够判断一张图片是否是猫的关联f。

学习就是需要从有限的例子中找到合理的f，那么一个方向1也就是训练更多的数据，看到更多的情况。比如有的同学就用题海战术。这就是这些年大数据对人工智能所起到的作用。

然而单靠这一个方向并不足够，第二个方向是那些只做一两道题就可以抓住问题核心的学霸们，而这个实际上才是深层学习在自然界的任务中超越其他机器学习的关键。也就是加入先验知识，调整假设孔空间

这里先要了解两个概念，为什么光靠大数据不够。原因在于符合训练集规律的f不只是一个，有很多种，比如我想要两个数字相加等于1，我可以让其中一个数字为1，另一个数字为0，也可以让其中一个数字为-299，另一个数字为300。都能完成任务。但是到底哪种方法在我以后的应用中合适。

可以看到上图中，通过学习得到拟合训练集数据的关联f，但是实际应用在测试集熵，那些没有见过的数据中，红色的部分却预测错了。我们实际是想要一个非常规整的螺旋形状。

比如高考，能够解题的方法有很多种。有的非常巧妙，但是这些巧妙的解题方式只适用于个别题目。再其他题目就不再适用。学生可能找到了一个可以解除所有他做过的题的解题方式，但却无法保证该解题方式在高考时同样有效。

只考虑离散的简单任务，1维的时候有10种情况，2维的时候100种，3维的时候有1000种。维度越高，我们越无法见到所有的情况。那么那些没见过的数据该如何预测？一般的方式就是取左右两个遇见过的数据的平均值，但是在高维情况下，这种做法非常不可靠。想要预测的可靠就需要更多的数据，并且是不一样的数据，这在现实中即便是大数据也难以实现。

但是如果仔细思考，变体到底是如何产生的，那么就有解决方法了。这也就是distributed representation的产生。比如各式各样的苹果就可以被拆分成三个因素来控制。假如每个因素有两种情况，那么我们只需要学习好这六种情况即可。并不需要见到8种变体。

而椭圆其实也有变体，什么样的形状是椭圆？我们可以依照同样的思路继续分解下去。关键在于将变体分解成因素，而因素若是变体时，则继续拆分。

但是这样的拆分方式是有前提的。因为这种方式其实是直接寻找数据是如何在自然界生成的。

如果在一个完全不同的物理规律的世界中，事物可能就不是以这种组合的方式形成，那么以这种拆分方式所学习到的关联f就无法很好的完成生成数据的工作。所得到的预测也不会可靠。

寻找合适的关联f好比是大海捞针。不同的机器学习算法只是更擅长在某个海域打捞而已。如果重点打捞某个区域，那么其他的区域就会被忽略。所以如果想要打捞的关联f可以存在于大海的任何地方，那么深层学习并不会比其他的机器学习算法更优秀。

但事实是自然界中的很多任务的关联f，并非像我之前所说可以出现在大海的任何位置。并非均匀分布，而是集中在特定海域，那些符合自然物理现象的特定海域。而深层学习就是擅长打捞该海域的机器学习算法。

但是到目前为止，所讲的仅仅还是浅层神经网络为什么优秀的原因。具有一个隐藏层的神经网络只要有足够的隐藏节点，就可以拟合任意的训练数据。但是，为什么深层学习比浅层学习要优秀。

以右图为例，每个图的左边是网络连接图，而右边是不同变体之间的关系图。链接在一切的圆圈表示一个节点的不同变体，分开的圆圈就是不同节点。如果像上半部分那样用浅层神经网络来学习，那么每个变体只被拆成了自己所需要的因素，并不会对其他的样本起到作用。

但如果用深层网络，就容易在a b c这三个因素的部分形成共享。

再拿编程做比较，我们不会去直接一步到位的编程，而是会分成很多个小模块。这些小模块可以在不同的需求下都适用，也就是共享。这样，就不需要每遇到一个细微变形和再重新编写一次。

但是可以使用深层的前提就是上一层的因素可以接着用于下一层的组成。想象如果一个网络可以和任何节点链接，那么这样的网络就没有任何的侧重。好比在你找人时被告知，它无处不在和他在所有地方。但是深层网络，这里特指是前馈神经网络并不是任意链接。它只会一层一层的形成，不同层之间不会有跳跃。并且形成的组合也是在同一层内进行的。这种组合方式恰好符合自然界的物理形成规律。所以深层学习并不会在所有任务当中都强大，但是在很多自然形成的数据当中就超越其他机器学习算法。

这时再回过头来思考生命与环境的关系，环境会变得越来越复杂，但是复杂是在原有因素的基础上增加依照物理规律形成的各种变体，而学习就是将这些变体再次变回因素，并且获得能够一个关联f1,这样不断的拆分，获得一连串的关联，

前馈神经网络可以适用于几乎所有的任务，但是它非常的一般性，所提供的先验知识的针对性很低。或者说并不足够。先验知识不够针对，那么训练所需要的数据量就会变大。并且过深之后会将那些噪音的形成规则也学习到模型当中，而这些规律并不是我们想要的。而神经网络的其他变体，比如循环神经网络，卷及神经网络就提供了更多针对的先验知识，可以缩小搜索的海域面积，排除掉那些噪音所带来的额外干扰。而神经网络的变体的设计也离不开共享二字。

关于深层的优势还有另外一个例子帮助理解。深层学习是以数据是如何生成的方式去寻找关联f。自然界的物质都并非从零开始重新生成，并不是说一个飞机是从头开始随机形成的，而是从原子，从分子，不断逐层形成。而形成飞机也好， 他们都有共用的元素，一百多个，又再次基础上形成分子。所以在学习飞机的原子是如何构造的同时，其实也等同于学习大炮的原子是如何构造的。这也就是为什么我们可以用深层神经网络进行迁移学习。

拆分因素：可以将变体拆分成因素附带关系的形式，降低训练所需数据量。

因素共享：使得所拆分的因素在不同的样本之间形成共享，可以用等量的数据量训练出更好的模型。

并不别扭，其实画面识别也好，语音识别也好。他根本不是在识别事物的真实状态。寻找我们人类是如何将物体的反光和声波的震动转换成相应概念的。比如中文发音会被关联到一种意思，英语发音会被关联到另一种概念的意思上。语音识别实际是寻找人脑是如何将震动关联到意思上的。画面识别也一样，人类也自己的关联方式，而如果要寻找狗的画面识别，所寻找到的关联将会非常不一样。所以画面识别，语音识别实际上是寻找人脑是如何将信息转变成概念的关联方式。你也可以理解成是如何加密的。

神经网络非常灵活，并没有任何固定的形式。不要认为循环神经网络就是循环神经网络，卷积神经网络就是卷积神经网络。如果你以这样的方式去学习神经网络，你这辈子都学习不完。因为网络的链接方式是无限的。你需要抓住至少两个核心。拆分因素，和因素共享。

比如因素拆分，

知道了深层神经网络的优势，那么如何设计神经网络也是基于这两点。前馈神经网络是基于这两点的最基本形式。

每层内的节点表示的是因素，这些因素共同描述事物的一种状态。这些状态可以逐层发展，发展后的状态又可以进行拆分和合并，来进行下一个状态的转变。

你可以把一个方框视为神经网络，并且神经网络可以继续与其他神经网络形成更深的神经网络。如右图，倒数第二个阶段的因素就是由三个神经网络提供的，不同神经网络的因素之间可以进行相加，也可以进行合并成更高维度的状态。 设计神经网络更像是玩乐高积木，但是玩的规则在于如何拆分因素，如何使不同的样本之间形成因素共享。所以在看到很多新的网络结构时，请务必考虑，他们的结构是如何考虑因素拆分和因素共享的。

下面就看看循环神经网络和卷积神经网络是如何利用因素共享这个优势来降低训练所需数据量和排除杂讯规律的干扰的。另外记住，卷积层并不是只针对画面识别，循环层也不是只针对时序信号。关键在于时间共享和空间共享

虽然大家习惯叫做循环神经网络，卷积神经网络，但是请以层的概念去理解它们。因为卷积神经网络处理完的输出可以接着进行其他处理，比如循环。这些都是状态的一种变化方式。

前馈层为了处理时序信号，需要将前后所发生的事物状态并成一个更大的向量。每一个维度都需要学习，即使很多维度根本都是无关紧要的。并且这样的方式只能处理固定长度的窗口。

但是如果告诉你，不同的时刻的状态都是由相同的规则进行处理的，那么一个时序信号序列有多个状态，每一时刻的状态都可以为循环层的权重的学习提供帮助。比如三个连续的时序状态在前馈层中

大家常说循环层比前馈层在处理时序

这种应用这种处理方式的原因是因为世界上很多的数据都符合时间共享的规律。

比如你在打击乐器，每个乐器的物理特性在时间维度上是不变的，并不会这个时刻是笛子的物理特性，而下一时刻变成了二胡的物理特性。

同时需要注意的是，循环层中有两个因素流，一个是从输入流得到的。在每个时刻共享。留一个信息流是从上一个时刻的状态得到的，同样是在时间维度上共享。所以你可以看到循环层有两个权重，两个

上述提到的是时间共享，还有空间共享的卷基层。前馈层同样可以进行画面识别，但是效果并不如卷基层给力。在tensorflow的教程中，如果用前馈层来训练，那么识别数字的准确率是92%，而用卷基层却可以达到97,8%的样子。关键在于加入空间共享这条先验知识后，降低了训练所需要的数据量。那么等量的数据就可以获得更好的学习效果。我说过画面识别实际是上是寻找人类是如何将反光关联到概念的。我们眼睛在观察时并非将整个图片全部收入眼帘。而是一个区域一个区域扫描。但是每个区域扫描的规则是共享的，你的眼球不会再扫描这个区域后在下个区域变成了猫眼睛。这样原本在

请回想无免费午餐定理，加入先验知识后。你实际上是调整了打捞关联f的区域，调整了假设空间。就拿线性方程y=ax+b来说，虽然你可以用y=ax的方式来把训练所需要的数据量从两个降低到1个，但是如果b根本就不等于零。那么只会让你寻找到的关联f变得不可靠

人们常说循环神经网络，RNN，比前馈神经网络要好的原因在于可以考虑到长关联性。但我曾经做过实验，用相同窗口长度的前馈神经网络和循环神经网络进行比较。循环神经网络依然比前馈网络效果要好。一定程度上验证了共享的作用。另外双向循环神经网络是两个循环层的堆叠，而堆叠后有两种合并方法，一种是相加，一种是合成并更长的向量。我做过的所有实验中，相加的效果都比合并要好。其实是因为第二个方向起到了限制作用，只有那些既可以在正向产生此种规律，又可以在反向产生另一种规律的关联f才会被我们采用

设计神经网络基本上就是在对假设空间进行调整

增加共享，降低确定关联f的范围所需要的数据量。如果用打鱼来比喻，增加共享实际上是减少了所需要的渔网的数量。

增加惩罚，我们可能会实现知道关联f或许不满足哪些特点，一旦这种特点的关联f产生时，就对优化算法得到的结果进行惩罚。有筛选作用。打鱼做比喻的话，那些不符合条件的小鱼会被大孔渔网筛选掉。

优化起点，就是我们

这些都是人类自身能力的延伸，这些技术可以应用在任何行业之中。

曾经我们是靠锄头耕地，现在靠机器，蹭我们靠算盘算账，现在靠计算机算账。但是机械也好，计算机也好，所更换的仅仅是工具，不是内容。我们始终需要耕地，需要算账

同样的人工智能仅仅帮助我们建立关联。有了合理的关联f，我们就可以根据特定情况产生相应的行为。 人工智能和机械和计算机一样，都是工具，并不会改变我们想要完成的任务。而未来的所有任务都可以依靠这一项新工具来实现。