AGI，从我做起

实现通用人工智能（AGI）道路之我见 III

前言

我对实现通用人工智能（AGI）有着非常浓厚的兴趣，平时经常思考一些实现AGI的最短路径，工作中也尽量选择一些能为AGI长期研究有所帮助的方向。但我深知自己能力有限，很多方面缺乏系统的理论指导。这次我把我的思考写下来，希望能够和大家一起讨论这个议题，得到各位专家的指正。

这一系列分为三篇文章："自上而下思考实现AGI技术难点及可能方法"，"稀疏是通往AGI的必由之路"，和"AGI，从我做起"。

这篇文章列举了一些我感兴趣，可以立刻着手去做的方向，也有一些我不明白，但对专家们来说可能很简单的方向。对于这些技术点我也希望能够找到一些业务方向能够立刻受益。希望大家可以私下和我联系，互通有无。

人们对AGI的认知

AGI是一个很模糊的词，没有一个确切的定义。每一个人的心中都有一个自己的AGI。有些人认为AGI并不存在（如Yann LeCun[1]），所有已知生物都不通用。这样讨论就转换为对生物缺陷的思考。

也有一些人从现状出发，思考AI具有什么样的能力就能够变得通用。这是一个移动的目标，现状的局限性限制了人们的想象力。从AI诞生开始这种趋势一直伴随着AI的成长。John McCarthy（AI的奠基人）曾经说过：As soon as it works, no one calls it AI anymore"[2]。

人们对AI的发展绝大多数采用自下而上的方法，渐变的改进现在的SOTA。研究方式是发散的，全方位的提高现今的科学技术水平。这种方式得到的进步当然非常坚实。但是缺点也比较明显，前进的速度很慢，有很多工作的effort对解决核心问题帮助不是很大，随着科技的整体进步有些工作的前提不复存在，导致工作的浪费。

我对AGI的认识如前文所述，自然也受现在发展阶段的局限，但我希望能够从自上而下的思考方式得到一些比较固定的关键技术目标，从而可以有计划的逐步攻克。在研究方式上希望能够跳跃性的以关键技术点为核心目标，研究成果辐射到其他非核心目标上。从而减少研究成本，加快进度。

AGI在学术界和工业界

在学术界和工业界直接以AGI为目标需要有很深的钱袋和长远的眼光。需要从算法、软件、硬件所有层面对现在的架构做出深度改变。

学术界总是更具前瞻性，但是现在的AI研究变成了金钱的游戏，有了钱，才能采集更多的数据，训练更大的模型。这对于资金缺乏、远离应用一线的教授们来说是比较困难的。这就需要从国家层面支持，比如DARPA提出了AI next campaign[3]，以$2B投资下一代AI实现contextual reasoning。虽然这只是实现AGI的下一步，一些关键技术可以在这一项目下得到研究。

工业界可以直接从AI的研究中得到收益，这些收益可以投资到更多的AI研究中去。因此现在AI的进展其实主要是由工业界驱动的。

工业界中有一部分是硬件公司，例如Intel，AMD，Nvidia。这些公司聚焦在设计高效执行已有AI算法的硬件。它们更像是一个跟随者，并不引领AI研究的潮流。

一些应用公司，例如Facebook，直接受益于AI算法的革新，有着很强的算法团队（FAIR）。但是它们的硬件研究团队比较薄弱，需要依赖于硬件公司提供支持。这种公司之间的隔阂阻止了一些革命性的研究。这些算法的研究被硬件束缚住了，不能完成研究螺旋式自主上升的闭环。

这种算法、软件、硬件都强的全站式公司放眼望去只有谷歌（Google）。谷歌算法研究世界一流（Google Brain），而且能够放长眼光，作出一些突破性架构探索，如depthwise separable convolution[4], transformers[5], mixture of experts[6]等等。谷歌软件作为发家老本行，自然没有什么说的。谷歌有世界上最顶尖的编译器专家，向MLIR，XLA，都是从谷歌提出的。而谷歌也是第一个看见ML硬件的重要性，早在2014年就开始设计专用加速硬件TPUv1/v2/v3。经过这么多年的迭代和经验积累，其硬件的水平早已甩出第二名几条街了。更可怕的是，谷歌本身也有广大的应用需求，在一家公司里打造了从应用、算法、软件、硬件全链路的闭环。这种迭代速度是其他公司不可比的。但是，谷歌也不是无懈可击。我认为，谷歌最开始设计TPU的架构选择systolic array对于当时稠密矩阵计算是无可厚非的最佳选择，但这种架构对于今后更有希望的稀疏架构就力不从心了，经过这么多年的迭代，谷歌也面临着船大难掉头的局面。这也是其他公司超越的机会。

DeepMind，作为谷歌下面一个独立分支，是现在公认AI研究最强的机构。有着谷歌硬件TPU的支持，DeepMind从事着一些最前沿的研究，从AlphaGo在围棋上一鸣惊人，到后来在Atari game超越人类，AlphaStar在策略游戏上占了上风，以及AlphaFold在预测基因序列上取得惊人成就。DeepMind也将实现AGI作为自己的长期目标[7]。DeepMind采用强化学习的方式（reinforcement learning）获得如此惊人成就。但在我看来，强化学习作为学习方法可以成为domain experts，但是要将复杂度更上一层楼，成为AGI，还是有所欠缺的。而且，reasoning，在我看来是一个比较简单的事情，但 是reasoning的基础一定要很有潜力，很坚实。现在基于convolution和稠密计算的transformer 能够提供给强化学习的输入还是有局限的。所以我们要集中精力在这些方面取得突破，然后强 化学习的能力自然就更上一层楼了。我提出的技术点和强化学习有些正交，攻克每一个技术点后都可以用强化学习的方法在应用中获得更好的体现。

AGI可以作为一个umbrella项目，一个长远目标，而其中的关键技术点攻关可以作为关键结果获取。同 时，这些关键技术点可以作为单独的项目，辐射到很多实际应用中去，带来实实在在的收益。而一部 分收益再反馈到下一些技术点攻关中。以这种方式，可以达到以小博大的效果。

我感兴趣的AGI近期可能项目分解

如前文所述，AGI的研究是一个算法、软件、硬件协同研究螺旋式上升的过程。算法在现有硬件的限制下作出先期突破，然后软硬件紧紧跟上，更有效的支持新的算法，从而使得算法能够在新硬件架构上能够得到进一步突破。

根据现阶段DNN发展的现状，我将现在没有依赖可以开始的项目分为关键技术点，可延展的项目，以及长期跟踪的项目。我希望各位能够引荐相应的专家一起讨论这些解决方案。

关键技术点：稀疏超级大模型

超大规模模型是现在DNN发展的一个重要方向。有着很重要的现实意义。我们现在已经能够训练千亿规模稠密模型（GPT-3），但到了万亿规模稠密训练就有点吃力，所以采用了粗粒度的稀疏MOE。我们训练十万亿或百万亿规模模型的时候，稠密训练是不可想象的。那时的问题不是要不要用稀疏，而是如何用稀疏。

有效执行完全稀疏的大模型需要硬件作较大的改动。因为硬件工程的开销很大，我们在开始硬件项目之前先要在算法和软件上对效果得到论证。

Transformer在self-attention上的稀疏

对权重的稀疏现在有了非常多的研究，对于激励细粒度稀疏的研究更多的是集中在ReLu引起的稀疏。从研究平衡的角度来说我们需要对其他激励稀疏的方式作深入探讨。而Transformer的self-attention的稀疏正好是非常重要的另一种稀疏方式。

Transformers是近一两年非常非常火的一类模型，从最开始在自然语言处理（NLP）发家，现在又渗入CV和推荐领域。大有一统天下的态势。Transformer的核心是它的self-attention机制，也引发了模型设计的另一个方向。

Transformer非常庞大。世界上最大的一些模型都是基于transformer架构，例如GPT-3，switch transformers。也是这种模型的诞生，对现在算力提出了数量级提升的要求。

Transformer的self-attention机制对于输入的计算量以sequence length的平方上涨，而long sequence输入的计算开销对于稠密计算是不可承受的。但是，我们已经非常清楚self-attention是非常稀疏的，需要有效的找到稀疏的方法来减少计算。现在已经有很多在这方面的研究了（就不举例了）。

因为Transformer模型非常流行，是很多应用场景的支柱。对这方面的算法、软件、硬件协同研究可以得到好的近期效益。

MOE稀疏

Mixture of experts（MOE）本身就是粗粒度的激励稀疏，有一个gating network根据输入在众多的experts中选择少数几个进行计算。Switch transformers就是对Transformer架构FFN采用MOE。这导致了很多系统的挑战。

因为模型巨大，训练需要采用数据并行、模型并行、流水并行。其中流水并行由于执行experts的机器有限，还有负载均衡的问题。现在的模型训练很多都采用非常大的batch size，MOE也不例外。这件事情本身是值得商讨的。如果你要在很多相同的硬件上做相同的处理，那大batch size可以 暴露更多的数据并行。但问题是，MOE是在不同的硬件上做不同的处理。在这种情况下，我们还需要这么大的batch size吗？

在这种情况下，是不是可以尝试在服务器间采用异步计算，但机器内采用同步计算呢？这是一种可以在现有硬件上实现的粗粒度稀疏。算法上可以采用同步SGD，但异步SGD性能可能更好。

对MOE的研究可以直接提高现有硬件对大型模型的执行效率。也可以研究一下软件的极限，为硬件 设计做准备。同时希望通过MOE的研究能够探讨一下在上文中提到的异步计算的潜力，从而为下一步稀疏研究打下基础。

异步训练软硬件协同设计探索

我在“稀疏是通往AGI的必由之路”一文中提到异步计算对稀疏训练的重要性。在算法上固然需要不停的寻找更好的不经过稠密训练步骤的稀疏训练算法，但同时在软硬件上也需要同时寻找更高效的异步稀疏训练的方式。

这一方面的研究可能稍微长期一点。异步编程模型在稀疏训练上的展现方式需要深度思考。软件框架需要隔离包装异步训练中通用部分，并且需要深度优化在现有硬件上的异步计算。这两方面取得一定进展后可以判定这个方向是否可以帮助已有业务。

同时，这些研究也可以了解现有硬件的瓶颈，指导硬件架构设计研究，对下一代硬件实现提出新的解决方案。

系统性探索这一研究所需时间会比较长。我们也可以针对一些具体问题作出一些点的突破，逐渐从点扩展到面，引起全面突破。

值得注意的是，同步计算总是可以实现异步计算的功能，差别是编程的effort和适用范围。这是考虑用同步还是异步的主要因素。另外，异步和同步的转换可以在不同层面进行，设计时也需要仔细思考。

关键技术点：记忆机制

记忆是时间上的稀疏，虽然我希望对空间上稀疏的研究能够对之有所借鉴，但是初期研究可能走的是不同的路，到后来我们再总结归纳也不迟。这样对记忆的研究初期也可以独立进行。

对记忆的研究有着显著的现实意义。现今的模型越来越大，为了使之不过拟合，需要更多的数据。所以现在ML的研究花了非常多的精力寻找更多的数据。supervised learning需要对每个数据加标注，非常费时费力，对验证集的精度要求又非常高。这些手工劳动开销非常大。self-supervised learning不需要加标注了，劳动力少一些，但大多用在预训练上。

对记忆的研究可以显著减少训练对数据的依赖性，本身就可以大幅减少数据采集的开销。对记忆的研究还可以实现few shot learning，显著减少训练开销。此外，对记忆的研究也会大幅改变现有模型架构，更综合发展DNN各个方面，防止DNN成为高智商的健忘儿。这些我在“稀疏是通往AGI的必由之路”有较详细的论述。

但是现在对记忆的研究还处于ad-hoc阶段，DeepMind研究过TM[8]和DNC[9]，但又没有了下文（我觉得这些研究过于模拟图灵机，有些不自然）。我总觉得这方面没有那么难吧，做一些入门级

的工作是不是就可以得到较好的结果呢？亦或这些研究需要做蛮多infrastructure工作，导致工作不多？

亦或已经有很多工作了，但我还不知道？我在这方面还是小白一个，希望有专家来指点迷津。

延展研究：稀疏联邦学习

我在“稀疏是通往AGI的必由之路”讨论过联邦学习。虽然联邦学习是一个较为独立的方向，但稀疏和联邦学习在一起有一些low hanging fruits，可以顺手摘了。此外，稀疏联邦学习和稀疏超大规模模型有些或隐或现的关系，说不定是远亲。主攻一方面但又不让另一面落后太多可以均衡的发展AI研究。

长期理论经验研究

在探索DNN未知领域时有两种方法，一种是根据经验尝试新的方法，得到新的结果。这种方法比较发散，不确定性高。现在DNN研究主要还是基于经验进行。另一种方法是以理论指导实践，找出问题的本质，在解决实际问题时只需要将问题映射到理论上就好了。但是现在理论很缺乏，对实际应用的指导较弱。现在的状况是理论追赶实践。

这两种方法会长期并存，我们在探索AGI的始终会遇到各种问题，对每一个问题都需要从实践和理论两方面着手。如果问题有一个较好的解决方案，我们需要立刻跳到下一个问题上去，而不让理论的缺失阻碍研究进度。

稀疏方法研究

稀疏的目的是只处理相关的权重和激励。但这相关不是非零即一，必定有一些弱相关的元素没有计算。和包括稀疏模型的稠密模型相比总是有精度损失。但是，如果和稀疏计算时间相当的稠密模型相比，稀疏的精度有所提高，稀疏计算的目的就达到了。

那在各种稀疏情况下（权重、激励、输出等），有没有什么理论和经验指导比较稀疏度和精度的关系。在没有稠密模型的情况下如何高效的找出最佳稀疏方式呢？这方面研究非常不足，虽然我们取得了一些初步进展[10]，但是还有很多未知领域需要探索。

异步SGD研究

当模型大到一定程度，同步SGD一定不是种有效的训练方式。我们现在已经对异步SGD有了很多研究，但是很少在实际应用场景上使用。而稀疏的异步SGD的研究进展我还不是特别了解。请熟悉这领域的人能够引我入门。

高阶理论探索

现在的DNN可以想象成将结果空间划分为非常多的小块，每一小块是输入的一个线性组合。我一直觉得这是一种复杂度非常高的方法。也很诧异这样的效果非常好。这连一个泰勒展开都不是！完全没有输入的高阶表达。我们为什么不能得出输入的二阶、三阶表达呢？

Transformer有一点点高阶的意思，但是中间用了softmax让我怀疑高阶并没有起作用。是不是这样呢？请理论大神给予指导。

Composition方法论

我在“自上而下思考实现AGI技术难点及可能方法”说明composition的重要性，也做了一些类比。这里就不重复了。但是，在DNN中采用什么样的composition的方法呢？如何找到composition的方法呢？我是完全没有想法。希望和各位同仁多多讨论，碰撞出一些火花。

总结

AGI是一个长期系统性的研究。其中每一个技术点都不是孤立或者只有理论意义的。每个技术点的攻克都可以辐射到很多有现实意义业务项目上去。虽然有些技术点很难，不知道怎么着手，但一旦成功了就会对AI的发展方向有着深远影响（我对此深信不疑）。

我花了一些时间，将我现在的思考写下来，组成三篇文章的一个系列：“自上而下思考实现AGI技术难点及可能方法”，“稀疏是通往AGI的必由之路”，和“AGI，从我做起”。希望能够抛砖引玉，和广大同行们一起讨论，加速研究，为早日实现AGI做贡献。

References

[1] https://twitter.com/ylecun/status/1204038764122632193
[2] https://cacm.acm.org/magazines/2012/1/144824-artificial-intelligence-past-and-future/fulltext
[3] https://www.darpa.mil/work-with-us/ai-next-campaign
[4] https://arxiv.org/pdf/1704.04861.pdf
[5] https://arxiv.org/pdf/1706.03762.pdf
[6] https://arxiv.org/pdf/1701.06538.pdf
[7] https://deepmind.com/about
[8] https://arxiv.org/pdf/1410.5401.pdf
[9] https://deepmind.com/blog/article/differentiable-neural-computers
[10] https://arxiv.org/abs/2106.09857