Hack VC合伙人：AI+Crypto的8大真正优势

作者：Ed Roman，Hack VC 管理合伙人

编译：1912212.eth，Foresight News

AI+ Crypto 是近期加密货币市场备受瞩目的前沿领域之一，比如去中心化 AI 训练、GPU DePINs 以及抗审查的 AI 模型等。

在这些令人眼花缭乱的进展背后，我们不禁要问：这究竟是真正的科技突破还是只是在蹭热点？本篇文章将为你拨开迷雾，剖析加密 x AI 设想和讨论其中真正的挑战和机遇，并揭示哪些是空洞的承诺，哪些又切实可行？

### 设想#1：去中心化 AI 训练

链上 AI 训练的问题在于需要 GPU 间的高速通信和协调，因为神经网络在训练时需要反向传播。 Nvidia 为此有两项创新（NVLink 和 InfiniBand）。这些技术使 GPU 通信变得超快，但它们仅限本地技术，仅适用于位于单个数据中心内的 GPU 集群（50+ 千兆位速度）。

如果引入去中心化网络，由于网络延迟和带宽增加，速度会突然变慢几个数量级。与从数据中心内的 Nvidia 高速互连获得的吞吐量相比，这种速度对于 AI 训练用例来说是不可能的。

请注意，以下也有创新可能会给未来带来希望：

训练的数据部分也具有挑战性。任何 AI 训练过程都涉及处理大量数据。通常，模型在具有高可扩展性和性能的中心化安全数据存储系统上进行训练。这需要传输和处理数 TB 的数据，并且这不是一次性周期。数据通常充满噪音并且包含错误，因此在训练模型之前必须对其进行清理，并转换为可用的格式。此阶段涉及标准化、过滤和处理缺失值的重复任务。这些都在去中心化环境中面临严峻挑战。

训练的数据部分也是迭代的，这与 Web3 并不兼容。OpenAI 经历了数千次迭代才取得了其成果。在 AI 团队中，数据科学家最基本的任务场景包括定义目标、准备数据、分析和整理数据以提取重要见解，并使其适合建模。然后，开发机器学习模型来解决定义的问题，并使用测试数据集验证其性能。这个过程是迭代的：如果当前模型表现不如预期，专家会返回到数据收集或模型训练阶段以改进结果。想象一下，这个过程如果在去中心化环境中进行，最先进的现有框架和工具在 Web3 中适配就变得不容易了。

另一个在链上训练 AI 模型的问题是，相较于推理，这个市场要无趣得多。目前，AI 大型语言模型的训练需要大量的 GPU 计算资源。从长远来看，推理将成为 GPU 的主要应用场景。试想为了满足全球需求，需要训练多少个 AI 大型语言模型，与使用这些模型的客户数量相比，哪个又更多？

### 设想#2：使用过度冗余的 AI 推理计算来达成共识

关于加密与 AI 的另一个挑战是验证 AI 推理的准确性，因为你无法完全信任单一的中心化方来执行推理操作，节点存在可能出现不当行为的潜在风险。这种挑战在 Web2 AI 中不存在，因为没有去中心化共识系统。

解决方案是冗余计算，让多个节点重复同样的 AI 推理操作，这样可以在无需信任的环境中运行，避免单点故障。

然而这种方法的问题在于，高端 AI 芯片极度短缺。高端 NVIDIA 芯片的等待时间长达数年导致价格上涨。如果你要求 AI 推理在多个节点上多次重新执行，那么就会成倍增加高昂成本，这对许多项目来说是不可行的。

### 设想#3：近期特定于 Web3 的 AI 用例

有人建议 Web3 应该有自己独特的、专门针对 Web3 客户的 AI 用例。这可以是（例如）使用 AI 对 DeFi 池进行风险评分的 Web3 协议、根据钱包历史记录为用户建议新协议的 Web3 钱包，或者使用 AI 控制非玩家角色的 Web3 游戏（ NPC）。

目前来看，这是一个初创市场（在短期内），其中使用案例仍在探索阶段。一些挑战包括：

长远来看，我们非常看好 Web3 原生的 AI 用例，特别是随着 AI 代理变得更加普遍。我们想象未来任何特定的 Web3 用户都会有大量的 AI 代理，来帮助自己完成任务。

### 设想#4：消费级 GPU DePIN

有许多依赖消费级 GPU 而不是数据中心的去中心化 AI 计算网络。消费类 GPU 非常适合低端 AI 推理任务或延迟、吞吐量和可靠性灵活的消费用例。但对于严肃的企业用例（这是重要的市场的大多数），与家用机器相比，客户需要更高可靠性的网络，并且如果他们有更复杂的推理任务，通常需要更高端的 GPU。数据中心更适合这些更有价值的客户用例。

请注意，我们认为消费级 GPU 适用于演示，以及能够容忍较低可靠性的个人和初创企业。但这些客户价值较低，因此我们认为专为 Web2 企业定制的 DePINs，长期来看将更有价值。因此，GPU DePIN 项目已经从早期主要使用消费级硬件，发展为具备 A100/H100 和集群级可用性的情况。

现实——加密货币 x AI 的实际用例

现在我们讨论能提供真正好处的用例。这些才是真正的胜利，加密货币 x AI 可以增加明显的价值。

### 真正好处#1：为 Web2 客户提供服务

麦肯锡估计，在分析的 63 个用例中，生成式 AI 每年可以增加相当于 2.6 万亿至 4.4 万亿美元的收入——相比之下，英国 2021 年 GDP 总额为 3.1 万亿美元。这将使 AI 的影响力增加 15% 至 40%。如果我们将生成式 AI 嵌入到目前用于用例之外的其他任务软件中的影响考虑在内，估计影响力大约会增加一倍。

如果你根据上述估计进行计算，这意味着全球 AI（超越生成 AI）的总市场价值可能达到数十万亿美元。相比之下，今天所有加密货币（包括比特币和所有山寨币）的总价值仅为 2.7 万亿美元左右。因此，让我们面对现实吧：短期内需要 AI 的绝大多数客户将是 Web2 客户，因为真正需要 AI 的 Web3 客户将只占这 2.7 万亿美元的一小部分（考虑到 BTC 是这个市场，比特币本身不需要 / 使用 AI）。

Web3 AI 用例才刚刚开始，目前还不清楚该市场规模有多大。但有一点是肯定的——可预见的未来，它仅占 Web2 市场中的一小部分。我们相信 Web3 AI 仍然有光明未来，但这仅仅意味着 Web3 AI 目前最强大的应用是服务 Web2 客户。

假设可以从 Web3 AI 中受益的 Web2 客户示例包括：

这是相对稳定的客户角色，因为客户通常规模大且有价值。他们不太可能很快倒闭，而且他们代表了 AI 服务的巨大潜在客户。为 Web2 客户提供服务的 Web3 AI 服务将受益于这些稳定的客户群。

但为什么 Web2 客户想要使用 Web3 堆栈呢？这篇文章的接下来部分阐述了这种情况。

### 真正好处#2：通过 GPU DePIN 降低 GPU 使用成本

GPU DePIN 聚合了未充分利用的 GPU 计算能力（其中最可靠的来自数据中心），并使其可用于 AI 推理。类比该问题的简单方法是「GPU 中的 Airbnb」。

我们对 GPU DePIN 感到兴奋的原因是，如上所述，NVIDIA 芯片短缺，而且目前有浪费的 GPU 周期可用于 AI 推理。这些硬件所有者付出沉没成本，并且目前没有充分利用设备，因此与现状相比，可以以低得多的成本提供这些部分 GPU ，因为这实际上为硬件所有者「找到了钱」。

示例包括：

注意，并非所有 GPU 硬件都适合 AI 推理。造成这种情况的一个明显原因是，较旧的 GPU 没有 LLMs 所需的 GPU 内存量，尽管已经有一些有趣的创新可以在这方面提供帮助。例如，Exabits 的技术可以将活动神经元加载到 GPU 内存中，将不活动神经元加载到 CPU 内存中。他们预测哪些神经元需要活跃 / 不活跃。这使得低端 GPU 能够处理 AI 工作负载，即使 GPU 内存有限。这有效地使低端 GPU 对于 AI 推理更加有用。

Web3 AI DePINs 需随着时间的推移发展其产品，并提供企业级服务，例如单点登录、SOC 2 合规性、服务级协议（SLA）等。这类似于当前云服务提供商为 Web2 客户提供的服务。

### 真正好处#3：抗审查的模型以避免 OpenAI 自我审查

关于 AI 审查制度的讨论很多。例如土耳其暂时禁止了 OpenAI（后来 OpenAI 提高了合规性，他们就改变了做法）。我们认为国家级的审查制度是无趣的，因为各国需要采用 AI 来保持竞争力。

OpenAI 也会进行自我审查。例如，OpenAI 不会处理 NSFW 内容。 OpenAI 也不会预测下一次总统选举。我们认为 AI 用例不仅有趣，而且市场巨大，但 OpenAI 出于政治原因不会触及该市场。

开源是个很好的解决方案，因为 Github 存储库不受股东或董事会的影响。 Venice.ai 就是一例，它承诺保护隐私并以抗审查的方式运营。 Web3 AI 可以有效地提升其水平，即在成本较低的 GPU 集群上为这些开源软件 (OSS) 模型提供支持，以执行推理。正是由于这些原因，我们相信 OSS + Web3 为抗审查的 AI 铺平道路的理想组合。

### 真正好处#4：避免向 OpenAI 发送个人身份信息

大型企业对其内部数据存在隐私担忧。对于这些客户来说，信任 OpenAI 第三方拥有这些数据可能很难。

Web3 中，对这些企业来说，他们的内部数据突然出现在去中心化网络上,可能看起来更加令人担忧（表面上）。然而，针对 AI 的隐私增强技术中存在创新：

可信执行环境（TEE），例如 Super Protocol

全同态加密 (FHE)，例如 Fhenix.io（Hack VC 管理的基金的投资组合公司）或 Inco Network（均由 Zama.ai 提供支持），以及 Bagel 的 PPML

这些技术仍在不断发展，并且通过即将推出的零知识 (ZK) 和 FHE ASIC，性能仍在不断提高。但长期目标是在微调模型时保护企业数据。随着这些协议的出现，Web3 可能会成为隐私保护 AI 计算更具吸引力的场所。

### 真正好处#5：利用开源模型的最新创新

过去几十年来，开源软件一直在蚕食专有软件的市场份额。我们将 LLM 视为某种专有软件形式，足以破坏 OSS。值得注意的挑战者例子包括 Llama、RWKV 和 Mistral.ai。随着时间的推移，这个列表无疑会不断增长（更全面的列表可以在 Openrouter.ai 上找到）。通过利用 Web3 AI（由 OSS 模型提供支持），人们可以利用这些新的创新来创新。

我们相信，随着时间的推移，开源的全球开发队伍与加密货币激励措施相结合，可以推动开源模型以及建立在其之上的代理和框架的快速创新。 AI 代理协议的一个例子是 Theoriq。 Theoriq 利用 OSS 模型创建一个可组合的 AI 代理互连网络，可以将其组装起来创建更高级别的 AI 解决方案。

我们之所以对此充满信心，是因为过去，随着时间的推移，大多数「开发者软件」的创新慢慢被 OSS 超越。微软曾经是一家专有软件公司，现在他们是为 Github 做出贡献的排名第一的公司。这是有原因的，如果你看看 Databricks、PostGresSQL、MongoDB 和其他公司如何颠覆专有数据库，那就是 OSS 颠覆整个行业的一个例子，所以这里的先例非常有说服力。

然而，这其中也有一个问题。开源大型语言模型（OSS LLMs）的一个棘手之处在于，OpenAI 已经开始与一些组织（如 Reddit 和《纽约时报》）签订付费数据许可协议。如果这一趋势继续下去，开源大型语言模型可能会因为获取数据的财务障碍而更难以竞争。Nvidia 可能会进一步加强对保密计算的投入，以作为安全数据共享的助力。时间将揭示这一切的发展。

### 真正好处#6：通过高削减成本的随机抽样或通过 ZK 证明达成共识

Web3 AI 推理的挑战之一是验证。假设验证者有机会欺骗他们的结果来赚取费用，因此验证推论是一项重要措施。请注意，这种作弊行为实际上尚未发生，因为 AI 推理还处于起步阶段，但除非采取措施抑制这种行为，否则这是不可避免的。

标准的 Web3 方法是让多个验证器重复相同的操作并比较结果。如前所述，这一问题面临的突出挑战是，由于目前高端 Nvidia 芯片的短缺，AI 推理的成本非常昂贵。鉴于 Web3 可以通过未充分利用的 GPU DePIN 提供更低成本的推理，冗余计算将严重削弱 Web3 的价值主张。

更有前途的解决方案是为链下 AI 推理计算执行 ZK 证明。在这种情况下，可以验证简洁的 ZK 证明，以确定模型是否经过正确训练，或者推理是否正确运行（称为 zkML）。例子包括 Modulus Labs 和 ZKonduit。由于 ZK 操作是计算密集型的，因此这些解决方案的性能仍处于初级阶段。不过，我们预计随着 ZK 硬件 ASIC 在不久的将来发布，情况会得到改善。

更有希望的是一种，有点「Optimistic」基于采样的 AI 推理方法设想。在这个模型中，只需验证验证者生成结果的一小部分即可，但将大幅削减的经济成本设置得足够高，这样如果被发现，就会对验证者的作弊产生强大的经济抑制作用。通过这种方式，你可以节省冗余计算。

另一个有前途的设想是水印和指纹解决方案，例如 Bagel Network 提出的解决方案。这类似于 Amazon Alexa 为其数百万台设备提供设备内 AI 模型质量保证的机制。

### 真正好处#7：通过 OSS 节省费用（OpenAI 的利润）

Web3 为 AI 带来的下一个机会是成本民主化。到目前为止，我们已经讨论了通过 DePIN 节省 GPU 成本。但 Web3 还提供了节省中心化 Web2 AI 服务利润率的机会（例如 OpenAI，截至撰写本文时，其年收入超过 10 亿美元）。这些成本节省来自于这样一个事实：使用 OSS 模型而不是专有模型来实现额外的节省，因为模型创建者并不试图盈利。

许多 OSS 模型将保持完全免费，从而为客户带来最佳的经济效益。但可能也有些 OSS 模型也在尝试这些货币化方法。考虑一下 Hugging Face 上所有模型中只有 4% 是由有预算来帮助补贴模型的公司训练的。其余 96% 的模型由社区训练。这个群体（96% 的 Hugging Face）具有基本的实际成本（包括计算成本和数据成本）。因此，这些模型将需要以某种方式货币化。

有一些提议可以实现开源软件模型的货币化。其中最有趣的之一是「初始模型发行」的概念，即将模型本身进行代币化，保留一部分代币给团队，将模型未来的一些收入流向代币持有者，尽管在这方面肯定存在一些法律和监管障碍。

其他 OSS 模型将尝试通过使用来货币化。请注意，如果这成为现实，OSS 模型可能会开始越来越类似于其 Web2 盈利模型。但实际上，市场将分为两部分，一些模型仍然完全免费。

### 真正好处#8：去中心化的数据源

AI 面临的最大挑战之一是寻找正确的数据来训练模型。我们之前提到去中心化 AI 训练有其挑战。但是使用去中心化网络来获取数据怎么样（然后可以将其用于其他地方的训练，甚至在传统的 Web2 场所）？

这正是像 Grass 这样的初创公司正在做的事情。 Grass 是一个由「数据抓取者」组成的去中心化网络，这些人将机器的闲置处理能力贡献给数据源，为 AI 模型的训练提供信息。假设，从规模上看，由于大型激励节点网络的强大力量，这种数据源可以优于任何一家公司的内部数据源工作。这不仅包括获取更多数据，还包括更频繁地获取数据，以使数据更加相关和最新。事实上，阻止去中心化的数据抓取大军也是不可能的，因为它们本质上是去中心化的，并且不驻留在单个 IP 地址内。他们还有个可以清理和标准化数据的网络，以便数据在被抓取后有用。

获得数据后，你还需要位置将其存储在链上，以及使用该数据生成的 LLMs。

注意，未来数据在 Web3 AI 中的作用可能会发生变化。如今，LLMs 的现状是使用数据预训练模型，并随着时间的推移使用更多数据对其进行完善。 然而，由于互联网上的数据是实时变化的，这些模型总是有点过时。 因此，LLM 推断的响应稍微不准确。

未来的发展方向可能是一种新范式——「实时」数据。这个概念是，当一个大型语言模型（LLM）被问到推理问题时，LLM 可以通过提示传输并注入数据，而这些数据是实时从互联网重收集的。这样，LLM 可以使用最新的数据。Grass 就在研究这部分内容。

特别致谢以下人士对本文的反馈和帮助：Albert Castellana、 Jasper Zhang、Vassilis Tziokas、 Bidhan Roy、Rezo, Vincent Weisser、Shashank Yadav、Ali Husain、Nukri Basharuli、Emad Mostaque、David Minarsch、Tommy Shaughnessy、Michael Heinrich、Keccak Wong、 Marc Weinstein、 Phillip Bonello、Jeff Amico、Ejaaz Ahamadeen、Evan Feng、 JW Wang。