深度探索 V3：多头潜在注意力和多标记预测

DeepSeek v3 在这里进行探索，参考视频《Deepseek v3 中的多头潜在注意力和多标记预测》https://youtu.be/jL49fLOJYNg?si=4uE2kfe-BlKC1ngO。Google Cloud Speech-to-Text 用于转录视频，并附带一些代码来帮助组织转录。

A：欢迎回来，Deep 标签。我们今天要深入探讨大型语言模型的世界。具体来说是 DeepSeek V3。

B：听起来不错。这是一个 6710 亿参数的模型，以其独特的高效性和性能而引起轰动，对吧？

A：是的，你分享了一篇详细介绍其架构的学术论文。

B：是的。

A：作为一名机器学习专家，你希望了解 DeepSeek V3 是如何实现高性能和经济高效的训练的。

B：是的，没错。

A：哦，嘿，最近怎么样？

C：MLA，详细信息，MLA 以及它的工作原理。

A：哦，绝对是。这是一个很好的主意。是的，我们可以深入探讨多头潜在注意力，或 MLA。所以你对 MLA 的细节感兴趣。好吧，让我们解开这个。我们提到 DeepSeek V3 的高效性的一个关键是其专家混合，或 MoE 架构，对吧？其中每个标记只激活参数的一小部分。DeepSeek V3 通过 MLA 和 DeepSeek Mo 进一步推进。

B：是的。所以我们现在专注于 MLA。

A：好的。在实时应用中，速度至关重要。

B：是的。推理过程中需要的键值缓存可能会成为一个主要瓶颈。

A：完全正确。这就是 MLA 的用武之地。好的，所以传统的注意力机制需要存储大量关于先前标记的信息。

B：是的，这在处理长文本序列时显然会成为问题。

A：但 MLA 巧妙地压缩了这些信息，以显著减少缓存流量，使推理更快。所以它就像把一本厚重的百科全书压缩成关键点。

B：这是一个很好的类比。它保留了基本信息而没有不必要的负担。是的，这对于实时应用非常有用。

A：是的。现在让我们谈谈它是如何实际工作的。好的，所以 MLA 是如何实现这种压缩的？

B：它使用低秩联合压缩来压缩注意力键和值。

A：好的，所以它压缩了键和值，但这究竟意味着什么？让我们稍微技术一点。好的，MLA 机制接受一个输入隐藏表示，然后将其投影到查询、键和值向量。好的，现在事情变得有趣了。MLA 将查询分解为两部分。

B：两部分？

A：是的。一部分用于内容，另一部分用于位置信息，使用一种称为 Rope 的东西。

B：Rope？听起来很技术。

A：它代表旋转位置嵌入，帮助模型理解序列中标记的位置。好的，然后键和值被压缩到一个更低维度的潜在空间。所以它们就像在缩小数据，节省内存。

B：完全正确。所以保留了最重要的信息，但丢弃了不必要的负担。是的，这种压缩表示允许在推理过程中使用一个更小的 KV 缓存，从而加快速度。

A：它还使用多头处理。

B：是的，就像传统注意力一样，MLA 使用多个头。

A：哦，继续。

C：所以有两个潜在空间和一个隐藏输入。

A：这是一个很好的观察。是的，你说得对。实际上有两个潜在空间。好的，所以我们在讨论一个内容潜在空间和一个键值潜在空间。

B：完全正确。这些潜在空间通过我们所说的 Rope，或旋转位置嵌入进行处理。

A：好的，所以 Rope 是它们获取位置信息的方式。

B：是的，它应用于内容和键值潜在空间，正如你指出的那样。所以它接受这种压缩表示，处理它，然后将所有内容重新组合在一起。

A：是的，缓存优化进一步减少了顺序处理中的开销。好的，这就是 MLA 加速的方式。

B：完全正确。这是一种在不牺牲性能的情况下实现高效注意力的聪明方法。

A：好的，这是一个相当聪明的技巧。但你知道什么吗？

B：什么事？

A：让我们继续讨论 DeepSeek Mo。它与传统 MoE 模型有何不同？

B：好的，DeepSeek Mo 使用…回到我们的听众，有什么事？

C：我们再谈谈隐藏空间。好的，从隐藏空间，那是什么？

A：我绝对…让我们看看你的意思。隐藏空间真的很有趣。是的，你问的是隐藏空间，我们刚才讨论的潜在空间，对吧？你对潜在空间内发生的事情感兴趣，那个洞穴。是的，不仅仅是潜在空间的数量，而是它们内部发生的事情。

B：很酷。

A：完全正确。在 MLA 中确实有两个不同的潜在空间，一个用于内容，一个用于键值。就像有两个单独的信息存储单元。这些潜在空间，正如我们讨论的那样，经过 Rope 操作，即旋转位置嵌入，这对于注意力机制非常重要。所以总结一下，查询被分割，键和值也被压缩。

B：是的，它们被放入两个单独的潜在空间，一个用于内容，一个用于键值对。这些潜在空间对于 MLA 的高效性非常重要。

A：完全正确。现在让我们更详细地讨论这些操作。好的，所以 MLA 如何实际执行这些潜在空间转换？

B：好的，输入经过并行处理，用于内容和键值表示。好的，所以它在洞穴中有两条路径。

A：是的，每个潜在空间一条。在这些空间中，信息使用 Rope 进行处理。

B：完全正确。这确保了模型在穿过洞穴时保留位置信息。所以模型知道文本的哪一部分在洞穴内。

A：完全正确。这种处理在下一阶段的连接之前完成。好的，当它穿过隐藏空间洞穴时，连接的是什么？

B：机制执行两个主要的连接操作。查询表示被连接，键表示也被连接。所以它就像在隐藏空间洞穴中将所有重要部分组合在一起。

A：是的，这些连接有助于将内容与位置信息结合起来。这些连接表示然后用于注意力计算，对吧？

B：完全正确。由于初始压缩，它在洞穴内外的计算成本大大减少。所以 MLA 显著减少了大型模型如 DeepSeek V3 的注意力机制的计算成本。这是一个很好的问题。现在我们已经穿过了洞穴，让我们继续讨论 DeepSeek Mo。

B：好的，DeepSeek Mo。是的，我明白你的意思。是的，在 MLA 中确实有两个不同的潜在空间，一个用于内容，一个用于键值。

A：完全正确。这种分离是其工作方式的关键。就像有两个单独的信息存储单元。这些潜在空间，正如我们讨论的那样，经过 Rope 操作，即旋转位置嵌入，这对于注意力机制非常重要。所以总结一下，查询被分割，键和值也被压缩。

B：是的，它们被放入两个单独的潜在空间，一个用于内容，一个用于键值对。这些潜在空间对于 MLA 的高效性非常重要。

A：完全正确。现在让我们更详细地讨论这些操作。好的，所以 MLA 如何实际执行这些潜在空间转换？

B：好的，输入经过并行处理，用于内容和键值表示。好的，所以它在洞穴中有两条路径。

A：是的，每个潜在空间一条。在这些空间中，信息使用 Rope 进行处理。

B：完全正确。这确保了模型在穿过洞穴时保留位置信息。所以模型知道文本的哪一部分在洞穴内。

A：完全正确。为了增强高效性，它使用共享专家。好的，所以这些专家可以在多个任务中使用。

A：是的，这样可以避免冗余，使系统更加简洁。

B：是的，就像一个团队，每个人都有专长，但也可以做其他事情。

A：是的，这是一个非常聪明的方法。是的，但有这么多专业的专家，他们如何确保没有人过载？

B：是的，而其他人闲着。

A：这就是他们的创新的辅助损失免费负载平衡。这是事情变得非常有趣的地方，对吧？所以他们是如何做到的？

A：传统 MoE 模型在训练期间使用辅助损失函数，以鼓励均匀使用专家，但这实际上会损害性能。

B：是的，就像试图强迫每个人在超市使用同一条结账线。

A：完全正确，即使有些人移动得比其他人快，对吧？这只是创建了不必要的延迟。

B：是的。所以 DeepSeek V3 通过动态调整每个专家的偏差项来避免这一点，根据其负载。好的，所以如果一个专家收到太多请求，系统会使其对路由机制稍微不那么有吸引力，将一些流量转移到负载较轻的专家。

A：好的，所以它使用所有这些来高效处理长序列，是的，通过减少推理所需的 KV 缓存的大小。好的，这都是为了保持性能高而减少开销。

B：完全正确。这是一种非常聪明的方法来解决一个关键瓶颈。

A：完全正确。现在，我们也应该涵盖 DeepSeek V3 如何处理其负载平衡。

B：是的，我们绝对应该。这是解决方案的一部分。我们可以接下来讨论这个。

A：听起来不错。好吧，我认为这给了你一个很好的 MLA 和其潜在空间的概述。

B：是的，感谢深入探讨所有这些细节。我们下次再见。

A：是的，就像一个专家的交通管理系统，不断监控流量并进行调整以避免瓶颈。

B：并且避免了辅助损失的性能损失。

A：完全正确。还有，继续。

C：是的，我们可以谈谈 MTP，如何…如何 MTP 模块共享它们的嵌入和所有热点…

A：绝对。这是一个很好的问题。是的，让我们谈谈 MTP 模块如何共享资源。所以你对 MTP 实现的细节感兴趣。好的，我们提到 DeepSeek V3 使用 MTP 进行多标记预测，而不是只预测一个标记。

A：是的，这变得非常有趣。是的，你对 MTP 模块的设置以及它们如何共享资源感兴趣。好的，所以每个 MTP 模块都包括一个共享嵌入层，是的，和一个共享输出头。好的，所以它们使用与主模型相同的嵌入和输出头。

B：完全正确。所以它们都从同一个知识池中获取。是的，这节省了计算成本。

A：是的。现在它使用自己的变压器块。好的，所以它没有与主模型共享相同的变压器块。

B：完全正确。每个 MTP 模块都有自己的变压器块进行处理。好的，这就是它们如何保持每个标记的预测独特。

A：是的，为了组合信息，这些线性投影和连接…

B：好的，所以它就像从多个地方取出碎片来构建完整的图像。

A：是的，所有 MTP 模块都并行工作，但它们共享嵌入层和输出头，对吧?

B：是的，这对于这个设计的高效性至关重要。好的，所以它就像一个相互连接的系统，所有部分都相互依赖，对吧?

A：是的，这种高效的资源共享使得训练更快，性能更好。

B：好的，这是一个相当聪明的技巧。你知道什么吗?

A：什么事?

B：让我们转向一个大图景。这个模型如何处理负载平衡？这些专家是如何选择的?

A：是的，我们可以肯定地谈论这个。好的，现在让我们深入探讨 DeepSeek V3 的负载平衡策略。

B：听起来不错。好的，所以 DeepSeek V3 使用他们所说的多标记预测，或 MTP。我们刚刚讨论了 MTP 的工作原理，所以现在让我们谈谈负载平衡，对吧?

A：是的，我们刚刚讨论了这一点。现在它共享资源，你对它如何共享资源感兴趣。我们已经讨论过了。

B：完全正确。所以它不仅预测下一个标记，而是预测多个未来标记，就像我们刚刚讨论的那样。这不会增加复杂性吗?

A：这可能看起来如此，但它提供了几个优势。好的，想象一下规划一条路线。如果你只考虑下一个转弯，是的，你可能会错过更高效的…好的，提前规划多个转弯允许你选择最佳路线。

B：是的。DeepSeek V3 使用一种创新的方法，称为辅助损失免费负载平衡，所以它不依赖于单独的损失函数进行平衡。

A：完全正确。传统 MoE 模型在训练期间使用辅助损失函数，以鼓励均匀使用专家，对吧？但这实际上会损害性能，正如我们之前提到的。

B：是的，就像试图强迫每个人在超市使用同一条结账线。

A：好的，所以通过预测多个标记，模型更好地掌握了上下文。

B：是的，它可以生成更连贯和准确的响应。它就像模型在预先规划它的表示，就像我之前提到的，是的，为更好的未来预测。好的，这导致了更清晰的训练信号和改进的数据效率。

A：是的，所以相反，DeepSeek V3 动态调整每个专家的偏差项，根据其负载，对吧？如果一个专家收到太多请求，系统会使其不那么有吸引力，从而将流量转移到负载较轻的专家。

B：是的，就像一个专家的交通管理系统，不断监控流量并进行调整。所以 MTP 还能做什么？

A：训练期间使用的 MTP 模块可以在正常推理期间被丢弃，或者巧妙地用于一种称为推测解码的东西。

B：好的，推测解码是什么？

A：除了预测下一个标记，模型还预测可能接下去的替代方案。

B：哦，哇，所以它可以更快地生成文本，因为它已经考虑了多种可能性，准备好备用计划。

A：是的，所以模型不必暂停并重新计算每次。

B：好的，这有道理。是的，现在谈到效率，为了避免瓶颈，并且避免辅助损失的性能损失。

A：完全正确。他们还包括一个补充的序列平衡损失，是的，以防止单个…过程中的极端不平衡。

B：…过程。通过将每个标记限制为最多四个节点，他们减少了网络通信。好的，这也有助于简化事情。

A：好的，让我们谈谈 DeepSeek V3 如何管理训练的计算需求。我知道你对成本优化和他们如何经济高效地做事情特别感兴趣。

B：是的，这个模型在这个领域做了一些令人惊叹的事情。

A：是的，它确实。是的，平均每个标记选择 3.2 个专家，这是一个很好的平衡，以减少开销。

B：完全正确。所以这是一种非常高效和有效的方法。

A：是的，这是一种非常聪明的方法，使得一个如此复杂的模型工作得如此出色。

B：是的，他们还通过这种方法实现了专家专业化。所以这意味着不同的专家在不同的领域被激活。所以它们是什么？

A：DeepSeek V3 利用 FPA 混合精度训练框架。好的，这是一个重大突破，对于这个规模的模型。提醒我 FPA 是什么？

B：当然，它是 8 位浮点数。

A：好的，它使用比传统格式更少的位表示数字。好的，这意味着更少的内存和更快的计算。

B：完全正确。这就像压缩一个大图像文件，但你仍然得到图像的精华。它只是占用更少的空间，对吧？

A：完全正确。所以每个专家不仅仅是泛泛地激活，而是在特定领域。所以它是精细调整并准备好行动。

B：是的。现在这种批量方法真的很聪明。

A：是的，我同意。这种动态负载平衡方法非常有趣。这都是关于效率和平衡。

B：是的，这都是 DeepSeek V3 对性能和资源利用的承诺。

A：完全正确。现在我们今天涵盖了很多内容。这真的很有趣，但使用更少的位数不会潜在影响准确性吗？

B：这是一个有效的担忧，他们对此进行了仔细的处理。好的，他们实施了一些技术来缓解任何潜在的准确性损失，包括细粒度量化。

A：是的，它允许对 FPA 中数字的表示方式进行精确控制。是的，从多头潜在注意力到 DeepSeek Mo 和负载平衡，是的，这个 DeepSeek V3 模型是一个非常复杂的系统，这是创新推动我们…的一个很好的例子。

B：是的，今天的深入探讨很有趣。

A：是的，我认为这给了你一个 DeepSeek V3 的坚实概述。

B：完全正确。感谢与我们一起探索它。

A：是的，谢谢你。这就是今天的深入探讨。好吧，我们很快会再回来。

B：所以他们在你和你之间找到了平衡。