**说 明 书**

**基于自适应跳层的大模型推理加速系统及方法**

**【专利申请日必须早于相关论文发表日(影响专利授权实质性因素)，如课题组存在与本申请技术相关之论文已见报(任何语种+任何平台)则请尽快与事务所联系，否则申请提交后将无法通过专利局审查】**

**【专利不保护应用于工程技术领域技术方案的推演过程】**

**专利法规定保护对象为“怎么做”的完整过程(具体实现本发明宣称技术效果的所有必要工程指导)。除上述外的“为什么”等内容(例如论文中记载的分析推理、可行性仿真验证等)不属于专利保护范畴，建议删除，如将发明操作和原理描述混淆将导致申请难以通过审查。**

**【组合多个现有技术于一体的发明创造或将其实现的过程，当任一部分均按现有方式工作且全文未揭示任何从未被公开的技术手段时，专利局将评价该申请不具有创造性。简单替换、拼凑现有技术或者实质上由不同发明创造特征、要素简单组合形成将被评价为非正常申请】**

**技术领域**

1. 本发明涉及的是一种大语言模型领域的技术，具体是一种基于自适应跳层的大模型推理加速系统及方法。

**背景技术【100字以内对现有技术缺陷的简述，不涉及本发明改进或效果】**

1. 大语言模型(LLM)的自回归特性让它在推理时具备上下文学习能力，但是却导致推理代价变得很高。跳层策略有望在不修改模型结构的情况下降低推理成本和延迟，具体方法是在推理时跳过那些不影响推理结果的层，现有方法包括早跳(early skipping)、周期跳过(periodic skipping)和早退(early exit)三个类别。然而，这些跳层策略有存在一定的局限性。
2. 经过对现有技术的检索发现，中国专利文献号CN118780372A公开日20241015，公开了一种进行模型剪枝无损加速大语言模型推理的方法及系统，属于人工智能技术领域，包括：利用遗传算法选择跳过层对LLM进行模型剪枝，通过在加速大型语言模型的推理过程中跳过LLM中的某些层进行加速，并生成推测Token序列；然后在验证阶段使用原始LLM进行一次性前向传递来验证这些推测Token序列，由验证通过的Token组成最终输出Token序列，确保最终输出与未更改的LLM生成的输出保持一致，从而在保持输出质量的同时实现推理加速。但该现有技术与本发明相比，其无法解决的技术问题包括……

**发明内容**

1. 本发明针对现有跳层策略需要额外微调(fine-tuning)、无法适配不同模型不同序列个性化跳层以及序列批处理困难的问题，提出一种基于自适应跳层的大模型推理加速系统及方法，能够自适应地选择要跳过的子层，最小程度影响模型的推理性能。通过在历史推理过程中高效学习重要性分布，可以构建预填充阶段(prefilling phase)的跳层策略；通过在动态中间数据中进行在线重要性学习，可以进一步在解码阶段(decoding phase)改进跳层决策。另外在批处理场景中根据不同子层的重要性分布特点，对不同的子层进行了不同的处理。本发明相比现有的跳层技术更加个性化，大大提高对不同模型和不同序列的适配程度，在对多个模型的测试中文本生成质量超过了现有的跳层技术。
2. 本发明是通过以下技术方案实现的：
3. 本发明涉及一种基于自适应跳层的大模型推理加速系统，包括：离线学习模块、在线学习模块和批处理模块，其中：离线学习模块收集并整合以往推理结果中预填充阶段的重要性信息，用于离线学习每个请求在预填充阶段待跳过的子层；在线学习模块收集并整合当前请求在解码阶段的少量内容，实时更新每个请求在解码阶段待跳过子层；批处理模块通过对待跳过的子层进行优化调整实现同时推理多个请求。
4. 所述的以往推理结果是指：
5. 所述的子层是指一个完整的transformer层中所包含的注意力子层和线性子层。它们通过残差连接，所以可以看成两个独立的子层。**【高亮部分歧义表示待核实】**
6. 所述的当前请求在解码阶段的少量内容是指：
7. 所述的离线学习模块包括：A单元、B单元、C单元以及D单元，其中：A单元根据...信息，进行...处理，得到...结果，B单元根据...信息，进行...处理，得到...结果，C单元根据...信息，进行...处理，得到...结果，D单元根据...信息，进行...处理，得到...结果。
8. 所述的在线学习模块包括：A单元、B单元、C单元以及D单元，其中：A单元根据...信息，进行...处理，得到...结果，B单元根据...信息，进行...处理，得到...结果，C单元根据...信息，进行...处理，得到...结果，D单元根据...信息，进行...处理，得到...结果。
9. 所述的批处理模块包括：A单元、B单元、C单元以及D单元，其中：A单元根据...信息，进行...处理，得到...结果，B单元根据...信息，进行...处理，得到...结果，C单元根据...信息，进行...处理，得到...结果，D单元根据...信息，进行...处理，得到...结果。
10. 本发明涉及一种基于上述系统的基于自适应跳层的大模型推理加速方法，在使用新模型进行推理时，完整执行前N个请求不进行跳层，计算其中每个子层的重要性，据此筛选出待跳过子层并计算待跳过子层的补偿方案，用于在跳过该子层时进行补偿；在在线解码时，完整执行前P个token不进行跳层，计算其中每个子层的重要性，通过阈值筛选出待跳过子层并计算待跳过子层的补偿方案，用于在跳过该子层时进行补偿；当多个请求确定待跳过的子层，通过合并多个请求的执行提高模型的吞吐量，其中：当解码阶段，每个请求有待跳过的注意力子层或线性子层为冲突层时，通过定制化GPU核函数实现部分请求执行或通过少数服从多数的原则判断是否跳过该子层。

**技术效果**

1. 本发明所揭示的从未被公开的技术手段为：通过离线学习模块和在线学习模块得到每个请求在预填充阶段和解码阶段待跳过的不重要子层，通过批处理模块让多个请求同时执行。其中注意力子层和线性子层被分开考虑，而不是将transformer层整体考虑，使得跳层策略更加灵活。离线学习模块决定了请求在预填充阶段的跳层策略，在线学习模块决定了请求在解码阶段的跳层策略。

**【基于已公开的技术手段进行参数调整或对象替换不属于本发明创新】**

1. 上述技术手段所带来的从未被公开的新功能/效果是：不需要额外算力微调模型、能够实现批处理执行请求，在相同加速比的情况下各项准确度指标要优于现阶段跳层模型。

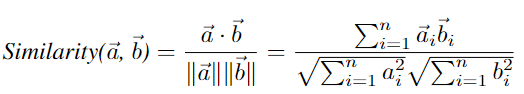
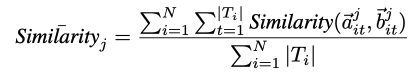
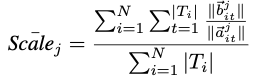
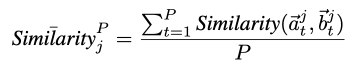
**【已公开技术手段产生的功能效果不属于本发明创新】**

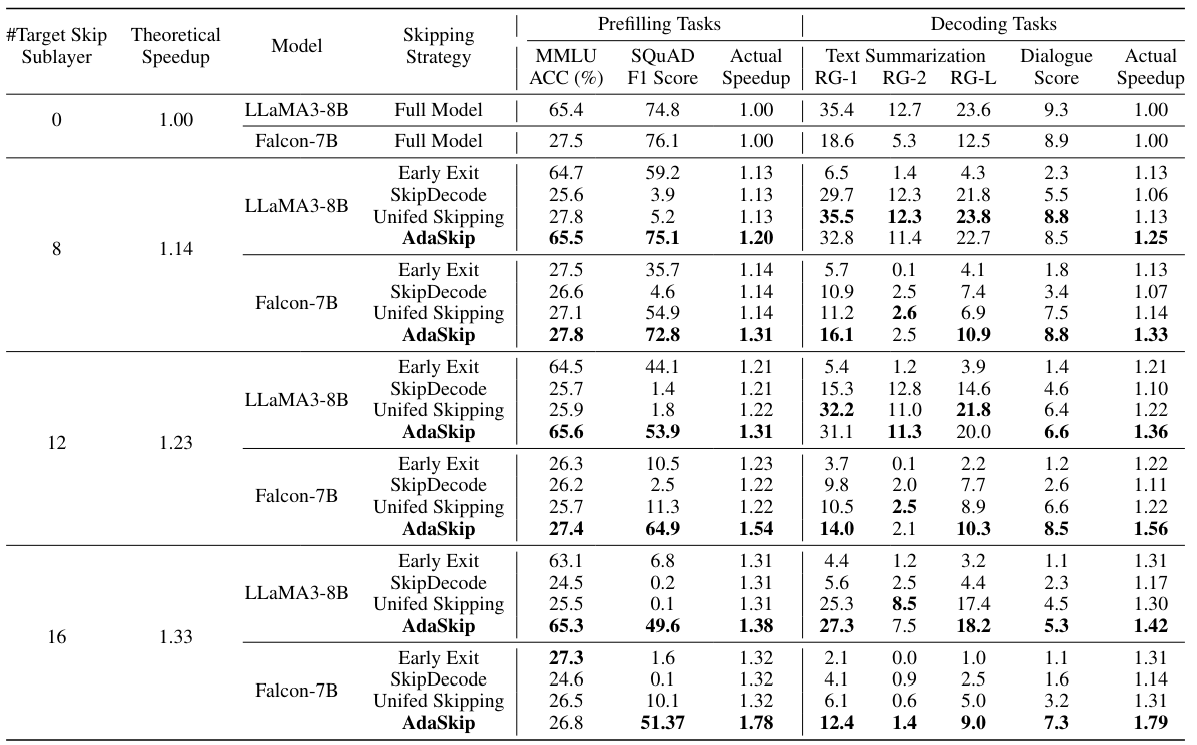
**附图说明**

1. 图1为本发明流程图；
2. 图2为本发明在两阶段的跳层示意图；
3. 图3为层重要性衡量指标图；
4. 图4为在线学习窗口长度的影响示意图；
5. 图中：(a)为注意力子层情况；(b)为线性子层情况。

**【建议提供至少三类说明书附图：(方法的)主流程图，(系统的)主拓扑模块结构示意图以及实施例效果示意图】**

**具体实施方式【待补充完整工程技术应用完整环节、实验数据、针对附图中技术内容的阐述；发明内容已出现的步骤操作不必重复】**

1. 如图1所示，为本实施例涉及一种基于自适应跳层的大模型推理加速方法，包括：
2. 第一步，对预填充阶段待跳过的子层进行建模：完整执行前N个请求不进行跳层，计算其中每个子层的重要性，据此筛选出待跳过子层并计算待跳过子层的补偿方案，用于在跳过该子层时进行补偿。**【所谓的“建模”得到何种具体模型？如未明确阐述模型的，此处建议删除建模或类似描述】**
3. 如图3所示，所述的重要性，即输入输出相似性，其中：a为某个子层的输入向量，b为输出向量。
4. 所述的筛选出待跳过子层是指：计算前N个请求在预填充阶段的层重要性统计值，根据加速比和每个层的重要性统计值，跳过m个最不重要的子层，具体为：第j层的重要性统计值，m的计算公式为？加速比是？
5. 所述的补偿方案，通过以下方式计算的得到：统计前N个请求在预填充阶段的输入输出差距，当跳过该子层后，使用输入输出差距统计值补偿输入，使得实际输出更接近理想输出，具体为：第j层的输入输出差距统计值，，其中：
6. 第二步，对解码阶段待跳过的子层进行建模：完整执行前P个token不进行跳层，计算其中每个子层的重要性，通过阈值筛选出待跳过子层并计算待跳过子层的补偿方案，用于在跳过该子层时进行补偿。**【所谓的“建模”得到何种具体模型？如未明确阐述模型的，此处建议删除建模或类似描述】**
7. 所述的阈值筛选是指：根据前P个token每个层重要性统计值，以已跳过的子层中的相似性最小值作为阈值，将超过阈值的层作为当前请求解码阶段需要额外跳过的子层，具体为：第j层的统计值为：，阈值，其中：
8. 所述的补偿方案**【与第一步相同？】**
9. 第三步，当多个请求确定待跳过的子层，通过合并多个请求的执行提高模型的吞吐量，其中：当解码阶段，每个请求有待跳过的注意力子层或线性子层为冲突层时，通过定制化GPU核函数实现部分请求执行或通过少数服从多数的原则判断是否跳过该子层。
10. 所述的冲突层是指：
11. 所述的定制化GPU核函数是指：
12. 所述的少数服从多数的原则是指：
13. 经过具体实验，在Pytorch 2.0, CUDA 12.2, NVDIA RTX 4090 GPU的环境设置下，按照P为20的配置运行上述方法，在多个模型上测试了多个数据集，多项结果实现现阶段最优，具体如下：
14. 如表1中间结果所示，展示了本发明在预填充阶段的效果。在目标跳过子层数量相同的情况下，本发明在MMLU和SQuAD任务中的表现都明显优于其他方法，这得益于本发明自适应地确定要跳过的最相似的子层。随着跳过的子层数量增加，本发明仍保持更高的生成质量，并且持续优于其他方法。另外，本发明在加速推理方面优势明显。因为本发明以子层的粒度进行操作，跳过了更加耗时的注意力子层，所以实际加速比相比于其他方法更优。
15. 如表1右侧结果所示，展示了本发明在解码阶段的效果。请注意，当跳过大量子层时，其他方法的生成质量会显著下降并失去可用性。相比之下，本发明仍然保持相对良好的生成质量，证明了其稳健性和有效性。另外，本发明在解码阶段的推理加速也十分明显。在给定相同生成质量的情况下，本发明比其他方法实现了更大的推理加速。
16. 表1



1. 为了证明本发明在在线学习窗口长度P设置为20的有效性，探索了不同P下预测层重要性的命中率。如图3所示，随着P的增大，命中率逐渐提高，这种改进是因为更长的窗口长度允许记录更接近当前上下文真实特征的特征。当P小于20时，命中率几乎随窗口长度线性增加；当P超过20时，命中率增长减慢，表明最佳窗口长度约为20。。
2. 与现有技术相比，本装置/方法的性能指标提升(即上述哪个环节采用了新技术得到的效果)在于：通过分析当前逐层跳过策略在加速推理执行方面的局限性，我们提出了一种新颖的自适应、逐子层的跳过系统和方法。本发明既无需训练，又适用于预填充和解码阶段。通过对各种数据集和模型进行严格测试，展示了本发明在生成质量和推理性能方面超越了现有的方法。
3. **上述具体实施可由本领域技术人员在不背离本发明原理和宗旨的前提下以不同的方式对其进行局部调整，本发明的保护范围以权利要求书为准且不由上述具体实施所限，在其范围内的各个实现方案均受本发明之约束。【专利制度基本原则在于“以对创新技术的充分公开换取垄断”，因此申请文件的主要目标在于客观清楚地表达发明人技术方案的具体创新改进点，由专利局本领域长期审查该领域技术的专业人员，就该改进点与现有技术进行客观比较并对存在创新的部分给予授权，如发明人未充分交底、故意隐藏关键技术、实验数据公开不充分或寄希望于审查阶段补充材料将直接导致申请文件无法实现上述目的】**

**说 明 书 附 图**

**根据专利法规定，附图仅为文字部分外的补充解释性内容，如附图中涉及具体的创新性改进则务必需在文字部分明示并详细阐明，仅记载在附图中的文字、字符、尺寸、结构等内容不视为专利保护对象**



图1

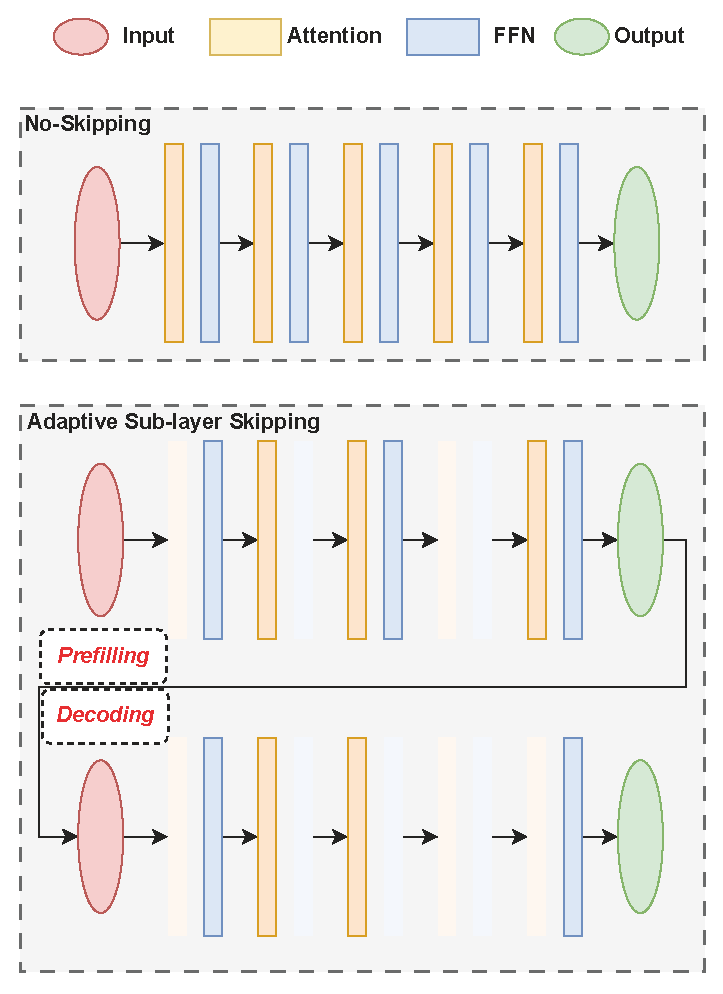


图2

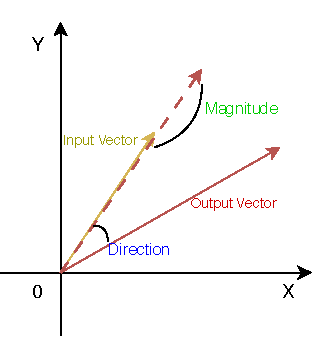
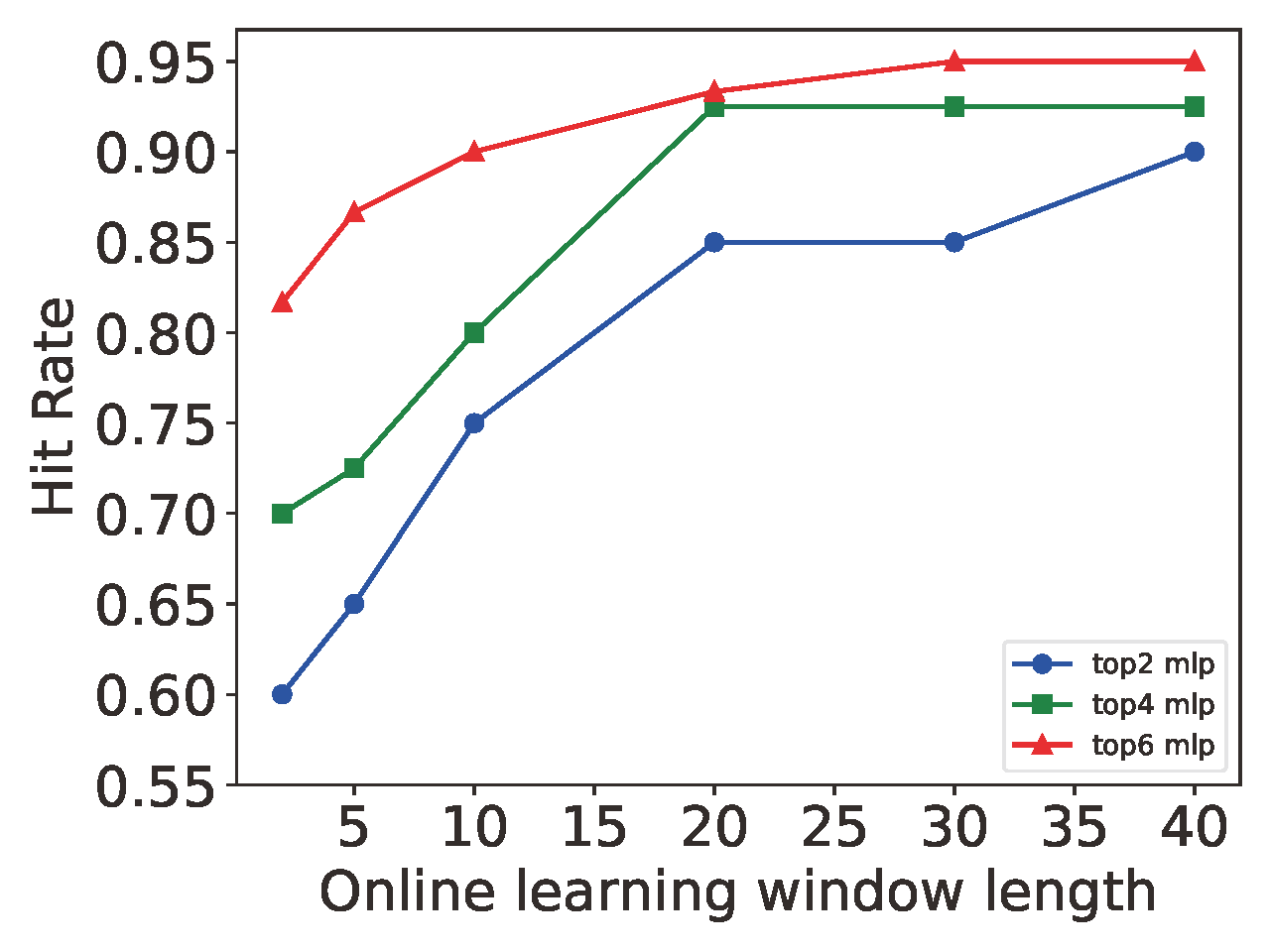
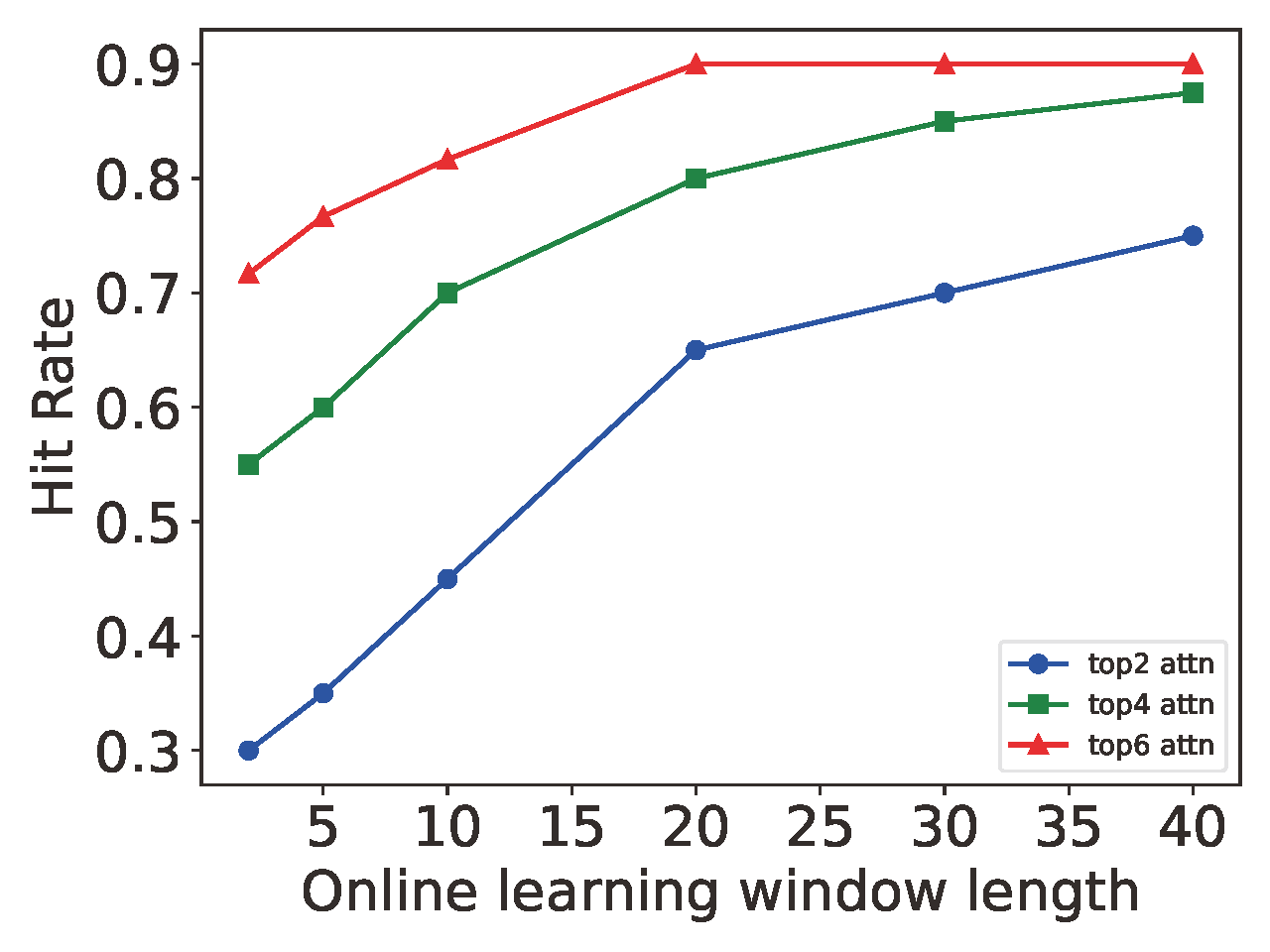


图3



1. (b)

图4

**权 利 要 求 书**

1、一种基于自适应跳层的大模型推理加速系统及方法，其特征在于，包括：离线学习模块、在线学习模块和批处理模块，其中：离线学习模块收集并整合以往推理结果中预填充阶段的重要性信息，用于离线学习每个请求在预填充阶段待跳过的子层；在线学习模块收集并整合当前请求在解码阶段的少量内容，实时更新每个请求在解码阶段待跳过子层；批处理模块通过对待跳过的子层进行优化调整实现同时推理多个请求。