**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】**

**【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**

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。

**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、序列批处理困难以及需要额外耗费算力对猜测的Token序列进行验证，当验证失败时需要回滚重算，且其无法节省KV Cache存储成本的不足，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：在无标签语料上使用大模型进行推理，收集推理过程中的层输入输出余弦相似度信息。
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，其顺序串联且内部都包含残差结构。
6. 所述的当前请求在解码阶段的少量内容是指：当前请求在解码阶段的前20步(20步为在线学习窗口长度)自回归解码，这部分被内容会被在线学习模块收集用于决定在20步后额外跳过哪些线性子层。
7. 所述的离线学习模块包括：注意力子层输入输出收集单元、注意力子层补偿单元、线形子层输入输出收集单元、线形子层补偿单元以及合并排序单元。其中：注意力子层输入输出收集单元会收集大模型在无标签语料库进行离线推理时产生的注意力子层输入向量和输出向量，计算输入向量和输出向量的余弦相似度；注意力子层补偿单元会计算注意力子层输入向量和输出向量的模长比值的平均值；线形子层输入输出收集单元会收集大模型在无标签语料库进行离线推理时产生的线形子层输入向量和输出向量，计算输入向量和输出向量的余弦相似度；线形子层补偿单元会计算线形子层输入向量和输出向量的模长比值的平均值；合并排序单元会将注意力子层与线性子层的输入输出向量余弦相似度进行综合排序，选出具有最高输入输出余弦相似度的子层，优先跳过。
8. 所述的在线学习模块包括：注意力子层输入输出收集单元、注意力子层补偿单元、线形子层输入输出收集单元、线形子层补偿单元以及额外跳层决策单元。其中：注意力子层输入输出收集单元会收集解码阶段前20步产生的暂未被跳过的注意力子层输入向量和输出向量，计算输入向量和输出向量的余弦相似度；注意力子层补偿单元会计算解码阶段前20步产生的暂未被跳过的注意力子层输入向量和输出向量的模长比值的平均值；线形子层输入输出收集单元会收集大模型在解码阶段前20步产生的暂未被跳过的线形子层输入向量和输出向量，计算输入向量和输出向量的余弦相似度；线形子层补偿单元会计算解码阶段前20步产生的暂未被跳过的线形子层输入向量和输出向量的模长比值的平均值；额外跳层决策单元会将暂未被跳过的子层的输入输出向量余弦相似度与离线阶段统计的相似度进行对比，额外跳过那些相似度超过离线阶段已决定跳过层中的最低相似度的子层。
9. 所述的批处理模块在批处理过程遇到多个请求跳层方案有差异时，通过少数服从多数的原则，使用CUDA编写的定制化GPU核函数来生成最终跳层方案。
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，当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。

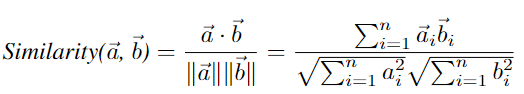
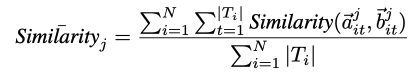
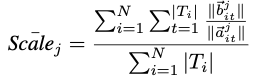
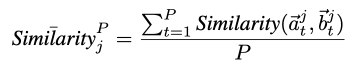
**技术效果**

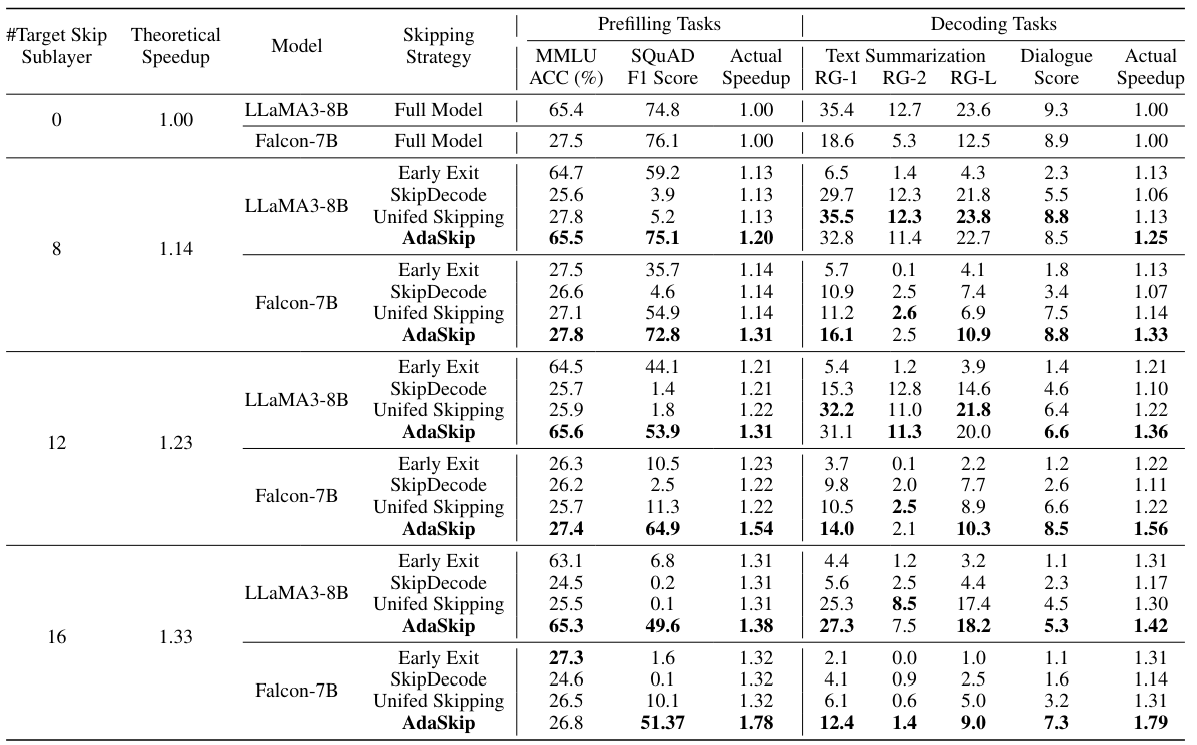
1. 本发明通过离线学习模块和在线学习模块得到每个请求在预填充阶段和解码阶段待跳过的不重要子层，通过批处理模块让多个请求同时执行。其中注意力子层和线性子层被分开考虑，而不是将transformer层整体考虑，使得跳层策略更加灵活。离线学习模块决定了请求在预填充阶段的跳层策略，在线学习模块决定了请求在解码阶段的跳层策略。与现有技术相比，本发明不需要额外算力微调模型、能够实现批处理执行请求，在相同加速比的情况下各项准确度指标要优于现阶段跳层模型。

**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。

**具体实施方式**

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，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。
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，预设参数m用于权衡加速比和精度，更高的m值会带来更高的加速比但也会导致更高的精度损失。在假设注意力子层和线形子层执行时间相同的情况下，加速比的计算方式为L/(L-m)，其中L为原始大模型总子层数。但由于实际执行中注意力子层与线形子层的执行时间不同，所以实际加速比需要测试。
5. 所述的补偿方案，通过以下方式计算的得到：统计前N个任务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：为第i个推理任务，为该任务的prompt长度，和分别为第i个推理任务中第j个token在第j层的输入输出向量。
6. 第二步，在在线解码阶段，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。
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：和分别为第i个推理任务中第j个token在第j层的输入输出向量。
8. 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，当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。
9. 所述的冲突层是指：某一子层对请求a而言需要跳过，而对请求b而言需要计算(不跳过)。
10. 所述的定制化GPU核函数是指：使用CUDA代码编写的、根据一个批(Batch)中每个请求的决策信息按照少数服从多数原则输出最终0/1决策(0代表跳过，1代表计算)。
11. 所述的少数服从多数的原则是指：当一个批(Batch)中超过一半请求决定跳过某一子层，则执行跳过，否则执行计算。
12. 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：
13. 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。
14. 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。
15. 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，本发明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。**

**说 明 书 附 图**



图1

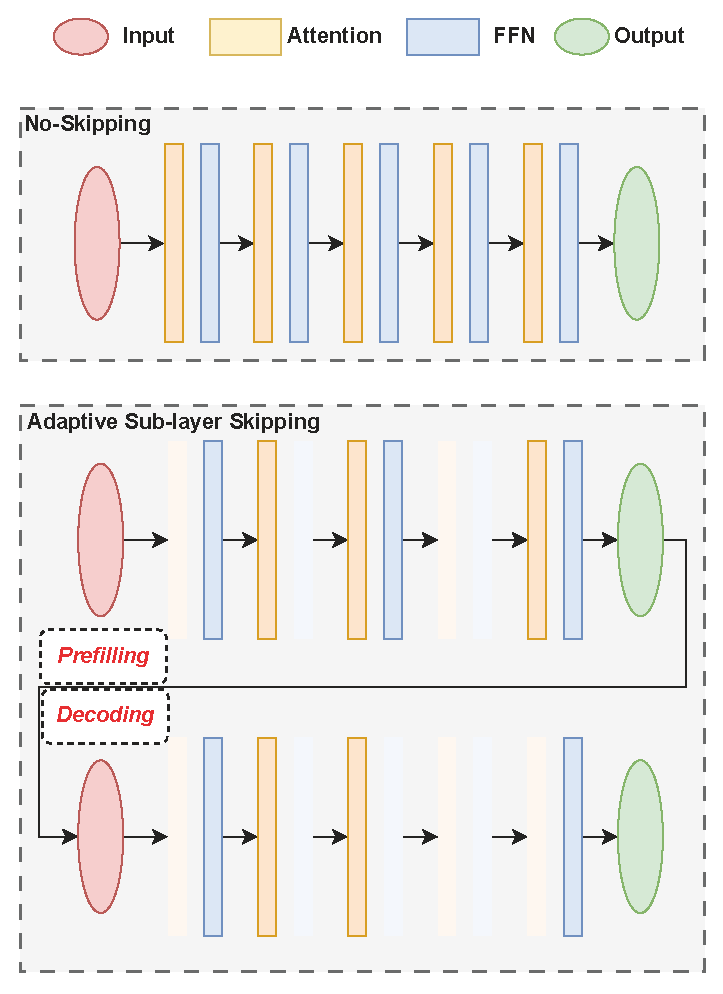


图2

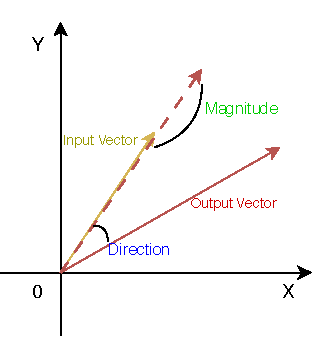
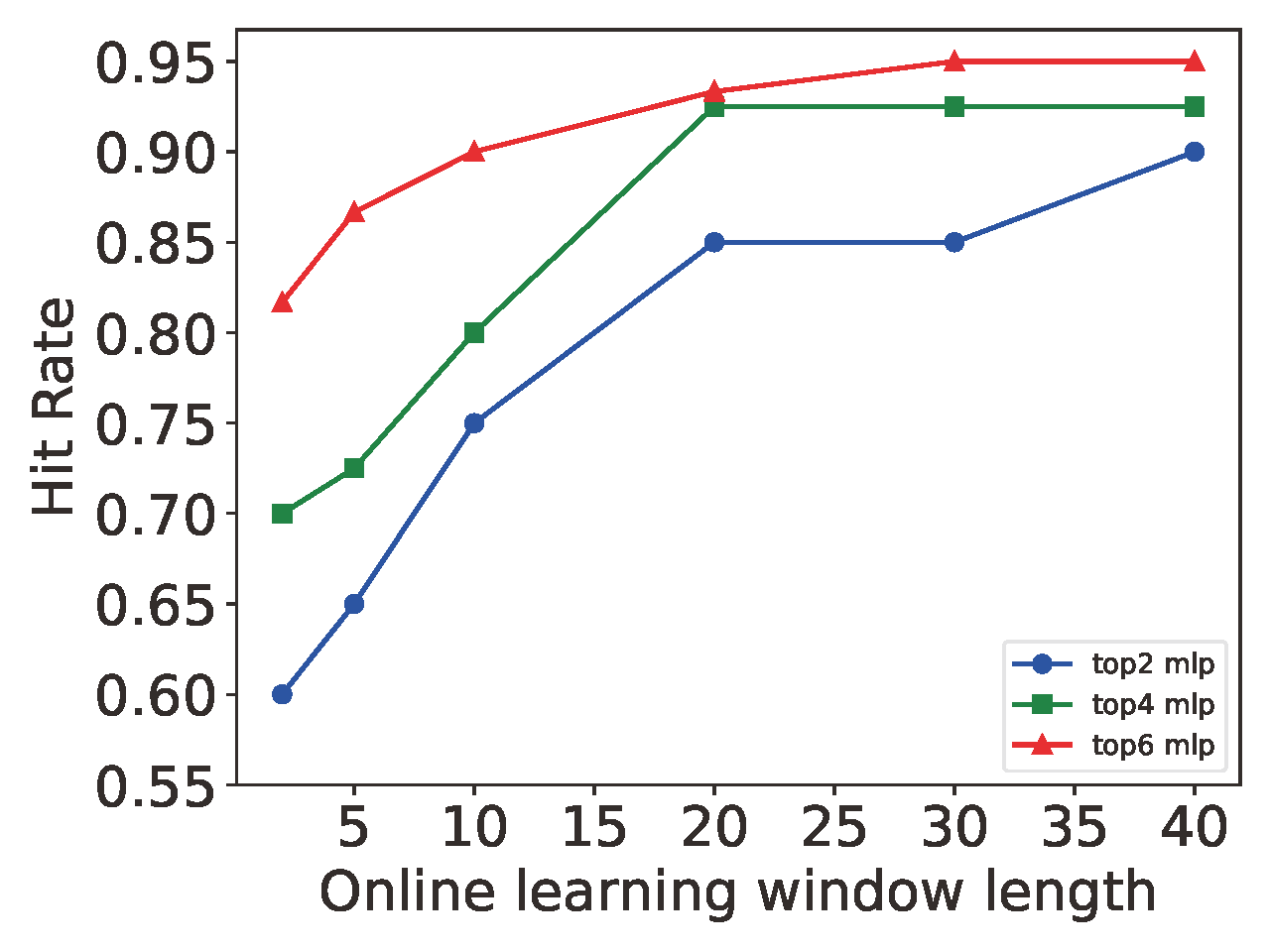
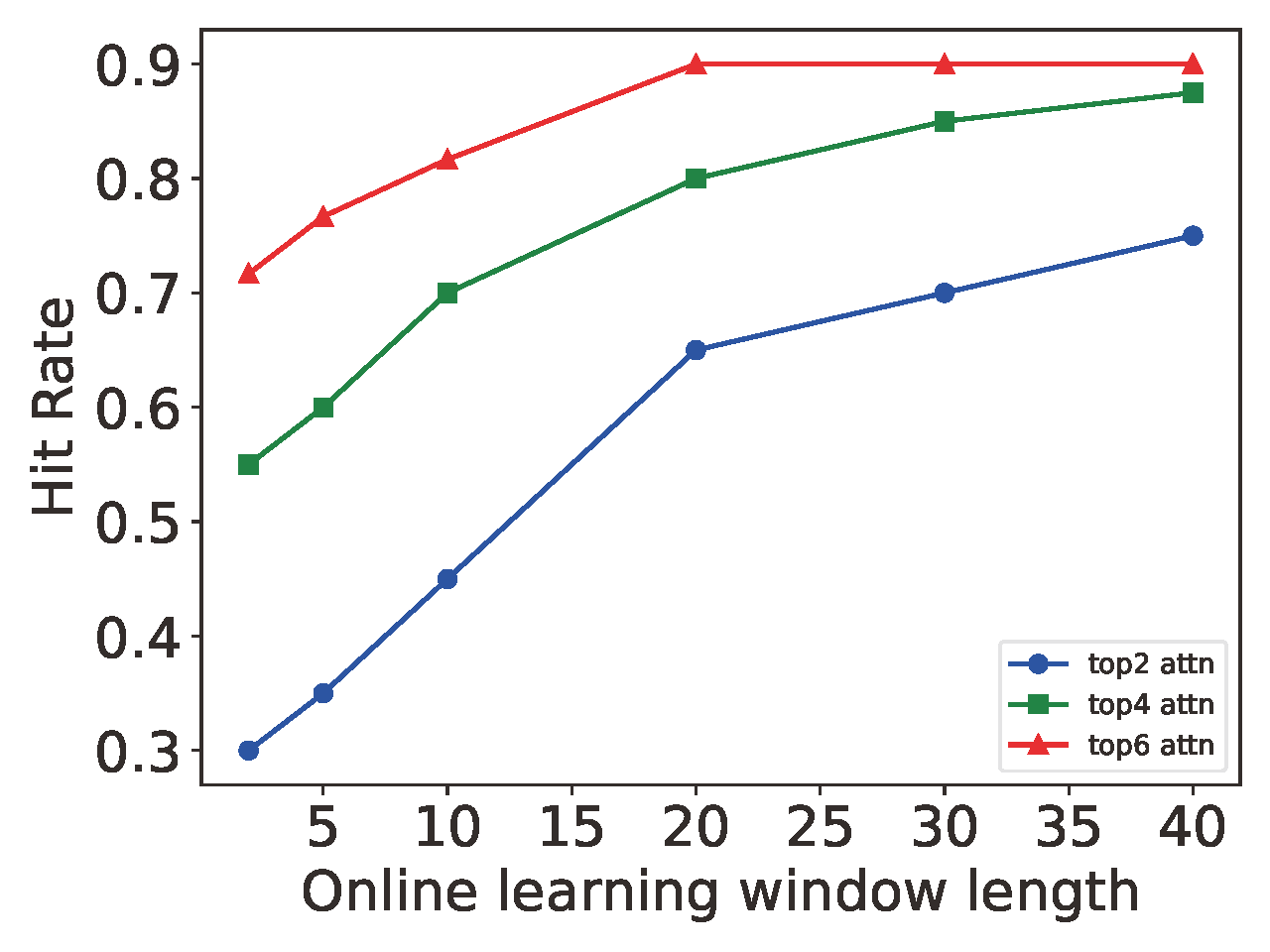


图3



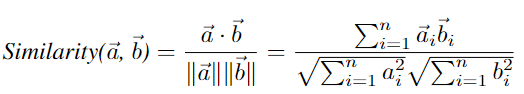
1. (b)

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，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；

所述的当前请求在解码阶段的少量内容是指：当前请求在解码阶段的前若干步自回归解码，以决定后续额外跳过的线性子层；

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。

2、根据权利要求1所述的基于自适应跳层的大模型推理加速系统，其特征是，所述的离线学习模块包括：注意力子层输入输出收集单元、注意力子层补偿单元、线形子层输入输出收集单元、线形子层补偿单元以及合并排序单元，其中：注意力子层输入输出收集单元会收集大模型在无标签语料库进行离线推理时产生的注意力子层输入向量和输出向量，计算输入向量和输出向量的余弦相似度；注意力子层补偿单元会计算注意力子层输入向量和输出向量的模长比值的平均值；线形子层输入输出收集单元会收集大模型在无标签语料库进行离线推理时产生的线形子层输入向量和输出向量，计算输入向量和输出向量的余弦相似度；线形子层补偿单元会计算线形子层输入向量和输出向量的模长比值的平均值；合并排序单元会将注意力子层与线性子层的输入输出向量余弦相似度进行综合排序，选出具有最高输入输出余弦相似度的子层，优先跳过。

3、根据权利要求1所述的基于自适应跳层的大模型推理加速系统，其特征是，所述的在线学习模块包括：注意力子层输入输出收集单元、注意力子层补偿单元、线形子层输入输出收集单元、线形子层补偿单元以及额外跳层决策单元，其中：注意力子层输入输出收集单元会收集解码阶段前若干步产生的暂未被跳过的注意力子层输入向量和输出向量，计算输入向量和输出向量的余弦相似度；注意力子层补偿单元会计算解码阶段前若干步产生的暂未被跳过的注意力子层输入向量和输出向量的模长比值的平均值；线形子层输入输出收集单元会收集大模型在解码阶段前若干步产生的暂未被跳过的线形子层输入向量和输出向量，计算输入向量和输出向量的余弦相似度；线形子层补偿单元会计算解码阶段前若干步产生的暂未被跳过的线形子层输入向量和输出向量的模长比值的平均值；额外跳层决策单元会将暂未被跳过的子层的输入输出向量余弦相似度与离线阶段统计的相似度进行对比，额外跳过那些相似度超过离线阶段已决定跳过层中的最低相似度的子层。

4、根据权利要求1所述的基于自适应跳层的大模型推理加速系统，其特征是，所述的批处理模块在批处理过程遇到多个请求跳层方案有差异时，通过少数服从多数的原则，使用CUDA编写的定制化GPU核函数来生成最终跳层方案。

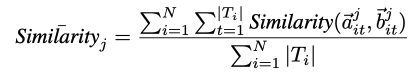
5、一种基于权利要求1-4中任一所述系统的基于自适应跳层的大模型推理加速方法，其特征在于，在使用新模型进行推理时，完整执行前N个请求不进行跳层，计算其中每个子层的重要性，据此筛选出待跳过子层并计算待跳过子层的补偿方案，用于在跳过该子层时进行补偿；在在线解码时，完整执行前P个token不进行跳层，计算其中每个子层的重要性，通过阈值筛选出待跳过子层并计算待跳过子层的补偿方案，用于在跳过该子层时进行补偿；当多个请求确定待跳过的子层，通过合并多个请求的执行提高模型的吞吐量，当每个请求有待跳过的注意力子层或线性子层为冲突层时，通过定制化GPU核函数实现部分请求执行或通过少数服从多数的原则判断是否跳过该子层。

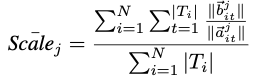
6、根据权利要求5所述的大模型推理加速方法，其特征是，具体包括：

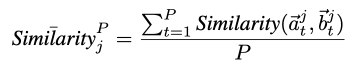
第一步，完整执行前N个请求不进行跳层，计算其中每个子层的重要性，据此筛选出待跳过子层并计算待跳过子层的补偿方案，用于在跳过该子层时进行补偿；

第二步，在在线解码阶段，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；

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，当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。

7、根据权利要求6所述的大模型推理加速方法，其特征是，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，预设参数m用于权衡加速比和精度，更高的m值会带来更高的加速比但也会导致更高的精度损失，在假设注意力子层和线形子层执行时间相同的情况下，加速比的计算方式为L/(L-m)，其中L为原始大模型总子层数，但由于实际执行中注意力子层与线形子层的执行时间不同，所以实际加速比需要测试。

8、根据权利要求6所述的大模型推理加速方法，其特征是，所述的补偿方案，通过以下方式计算的得到：统计前N个任务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：为第i个推理任务，为该任务的prompt长度，和分别为第i个推理任务中第j个token在第j层的输入输出向量。

9、根据权利要求5所述的大模型推理加速方法，其特征是，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：和分别为第i个推理任务中第j个token在第j层的输入输出向量。

10、根据权利要求5所述的大模型推理加速方法，其特征是，所述的定制化GPU核函数是指：使用CUDA代码编写的、根据一个批(Batch)中每个请求的决策信息按照少数服从多数原则输出最终0/1决策(0代表跳过，1代表计算)。

**说 明 书 摘 要**

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。

**摘 要 附 图**

