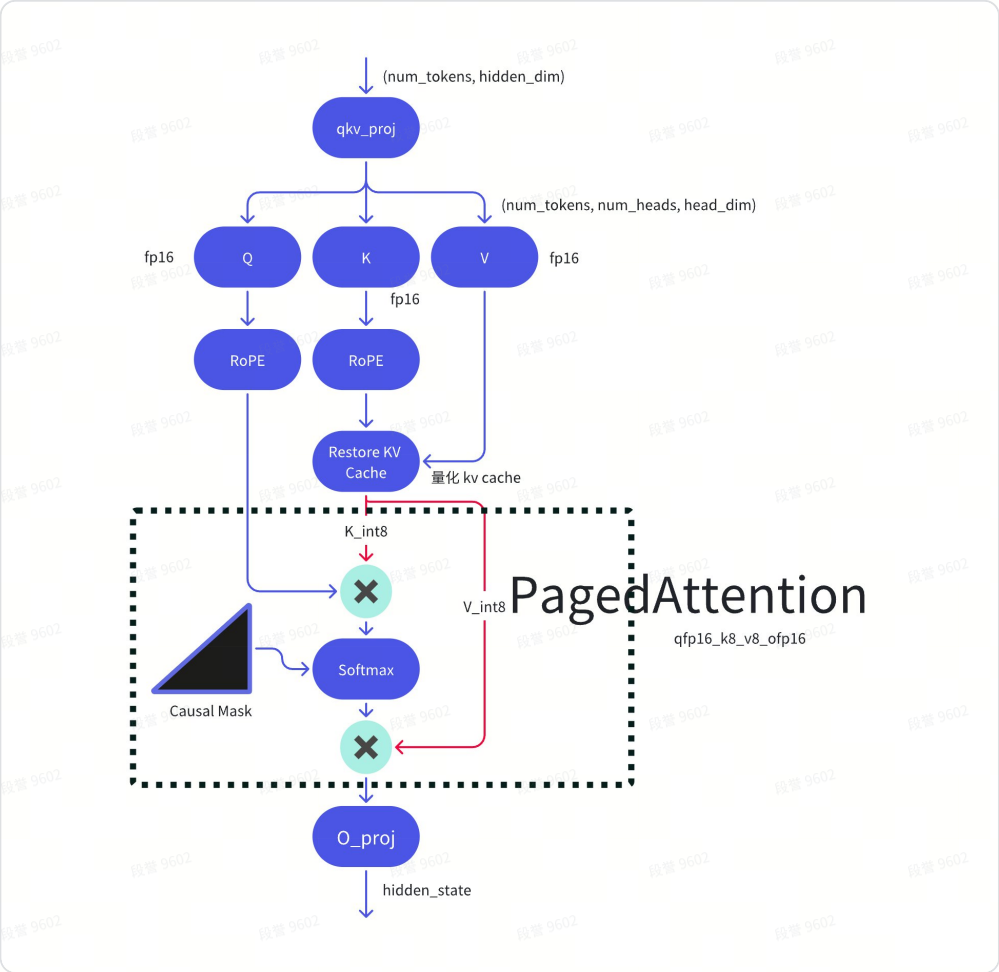
KV cache 量化

建议先阅读文档： 量化推理

量化会对精度造成一定影响，per token 相比 per tensor精度更好，但是性能上可能差强人意（原因见上述链接文档）。

LLM 推理中，运行时，绝大部份显存消耗都在 kv cache 上，因此，只对 kv cache 量化（类似weight only），理论上也有将最大batch 提升一倍。

计算流程图



如图中所示，PagedAttention 的输入 Q 为 fp16， K和V 为int8， 在 PagedAttention 算子内部，load int8 的kv值，然后反量化为 fp16，再进行 qkv 的计算。

PagedAttention 算子理论上受限于 memory bind，因此 load int8 的 kv cache，一定程度上可以提高算子性能。

所需的推理上的主要工作量为：

1. 在 Restore KV Cache 算子中增加量化代码，该算子需要量化 kv 值

2. PagedAttention 修改（重点）
3. 组网和其他 api 修改

精度影响：

理论上，llm 中的 gemm 都没有量化，仅仅量化 kv cache，其他都保持不变，对精度影响很小

重点：这里只做 kv cache 的量化，其他如权重，其他中间激活值，gemm 等均未量化

精度初步验证

这里的 scale 采样对称量化，**运行时动态计算（只考虑prompt kv的分布）**，这对首 token 有一定影响，后续为提前计算好 scale。

输入几个不同的 prompt，对比 bf16 和 int8 的输出，使用和线上一样的采样参数：

top_p = 1, top_k = 3, temperature=0.66606945, repetition_penalty = 1.02

prompt	**bfloat16**	**int8**	备注
[SEP][USER]你叫什么名字[SEP][BOT]	' 我是天工AI助手，很高兴为您服务。 '	' 我是一个人工智能助手，我可以帮助你回答问题和提供信息。\n '	没有说出名字
[SEP][USER]周亚辉是谁？[SEP][BOT]	' 周亚辉是中国知名的企业家和投资家，出生于1977年12月，四川省成都市人。他是昆仑万维公司的创始人、董事长，也是中国最大的动漫游戏公司之一。\n '	' 周亚辉是中国的一位企业家和投资家，他是昆仑万维的创始人和CEO，也是中国互联网公司在海外上市的第一人。他在创业和投资领域有着丰富的经验和成功案例。\n	
[SEP][USER]窗前明月光下一句是什么[SEP][BOT]	' 疑天高一尺。 '	' 疑天高一尺。 '	回答错误
[SEP][USER]如何看待中美关系[SEP][BOT]	' 中美关系是世界上最重要的双边关系之一，对全球和平、稳定与繁荣具有深远影响。以下是对中美关系的一些看法：\n\n1. 复杂性与多维性：中美关系是复杂且多维的，涵盖了政治、经济、安全、文化等多个领域。这种关系的复杂性要求双方以开放、包容和合作的态度来处理分歧，寻找共同利益和合作点。\n2. 竞争与合作并存：中美两国在许多领域存在竞争，但同时也存在广泛的合作空间。竞争可以激发两国的创新	' 中美关系是世界上最重要的双边关系之一，对全球和平、稳定与繁荣具有深远影响。以下是对中美关系的一些看法：\n\n1. 复杂性与多维性：中美关系是复杂且多维的，涵盖了政治、经济、安全、文化等多个层面。这种复杂性使得双方	

	精神和进取心，而合作则有助于解决全球性挑战，如气候变化、公共卫生危机等。\n3. 相互依存的经济关系：中美两国在经济上高度相互依存。贸易、投资和技术合作是中美经济关系的重要组成部分。这种经济上的相互依存为两国关系的发展提供了巨大的机遇，同时也带来了挑战。\n4. 需要管控分歧，避免冲突：中美之间存在一些敏感和复杂的问题，如台湾问题、南海问题等。双方应通过对话和协商，寻求妥善解决这些问题的方法，避免采取可能引发冲突的行动。\n5. 促进人文交流与合作：人文交流是增进中美两国人民相互了解和友谊的重要途径。双方应加强教育、科技、文化等领域的交流与合作，为两国关系的发展注入更多的人文内涵。\n6. 应对全球性挑战：在全球化时代，中美两国面临许多共同的全球性挑战，如气候变化、公共卫生危机、恐怖主义等。双方应加强合作，共同应对这些挑战，维护国际秩序和世界和平。\n\n总之，中美关系是一种重要且复杂的双边关系，需要双方以开放、包容和合作的态度来处理。只有通过加强沟通、增进理解、扩大合作，才能推动中美关系持续健康稳定发展。'	需要不断增进理解、加强对话，以求妥善处理分歧和摩擦。\n2. 合作潜力巨大：中美两国在许多领域都存在巨大的合作潜力。例如，在应对气候变化、公共卫生危机、恐怖主义等全球性挑战方面，中美合作具有至关重要的意义。此外，在经贸、科技、教育、文化等领域，双方也有广阔的合作空间。\n3. 竞争不可避免：随着中美实力的不断变化，竞争在两国关系中的成分也在增加。这种竞争既包括经济竞争，也包括科技、军事等领域的竞争。然而，竞争不应是零和的，而应是双方通过不断创新和进步，实现互利共赢。\n4. 合作有利于双方：中美之间的合作不仅有利于两国的发展，也有利于全球的繁荣和稳定。双方在应对全球性挑战方面的合作，有助于维护国际秩序和多边主义，推动构建人类命运共同体。\n5. 管控分歧的重要性：中美之间存在一些历史遗留问题和现实分歧，这需要双方以建设性方式加以管控。通过对话、协商和互谅互让，双方可以找到解决问题的途径，减少误解和误判，防止冲突升级。\n6. 推动关系持续改善：为了推动中美关系持续改善，双方需要采取积极措施，包括增进战略互信、加强各层级交流、推动经贸合作、管控分歧等。同时，双方还应尊重彼此的社会制度和发展道路，减少对彼此主权和领土完整的干涉，为两国关系的健康稳定发展创造有利条件。'	

精度看起来还行！

最大batch

以 13B llama 模型为例，使用 vllm 推理时：

Kv cache blocks number	bflaot16	int8
A100	3078	6127

能够存取 kv cache 的 block 数量翻倍，因此理论最大 batch 翻倍。

性能/吞吐测试

TODO

附录：PagedAttention

这部分一起详细剖析下 pageattention 算子的算法设计：