# 通过迭代精炼思想增强句子嵌入：一种改进文本理解的新方法

**摘要：** 本文提出了一种新颖的迭代方法——新精炼思想（Refined Thought）方案，旨在通过使大型语言模型执行受控的多通道信息聚合来提高句子嵌入的质量。该方案的核心原则包括：从基座 Embedding 模型的 last hidden layer output 中生成一个注意力引导的“记忆标记”，策略性地放置该标记以实现最佳信息融合，以及在迭代过程中精细管理位置编码。新 RT 方案以循环处理、记忆增强和注意力机制等既有理论为基础，旨在促进更深层次的文本理解。通过在 PAWSX、STSB 和 STS_Algorithm_jd2cv 等多种语义任务上的广泛实验验证，结果表明新 RT 方案能够持续提升性能，尤其是在复杂任务上，同时保持基座模型表征空间的完整性。本研究结果证实，这种有原则的迭代精炼能够产生更鲁棒、更高质量的句子嵌入，在文本表征学习方面取得了显著进展。

## 1. 引言

高质量的句子嵌入是众多下游自然语言处理（NLP）任务的基础，包括语义搜索、文本分类和问答系统。它们作为文本的密集、有意义的表示，使模型能够捕捉语义相似性和上下文细微差别。随着对日益复杂文本理解的需求增长，需要超越单通道处理的方法来提取更丰富、更精细的表示。

迭代处理，即模型多次重新审视其输入或中间表示，已显示出增强模型能力的潜力，通过允许“更深入的思考”或精炼来提升性能。新精炼思想（RT）方案提出了一种受控、有原则的方法，用于这种多通道信息聚合，以提高从预训练语言模型中获得的句子嵌入的质量。

新 RT 方案旨在通过在每次前向传播中生成一个单一的、注意力引导的“记忆标记”，并使其影响随后的前向传播，从而促进这种“进一步思考”。这个过程旨在通过允许模型迭代地整合信息来精炼最终的句子嵌入。与那些鲁棒性较差的迭代尝试不同，新 RT 方案优先保持基座模型固有的特性，例如其单向注意力机制和位置编码逻辑，从而确保稳定性和避免表征空间退化。

本文结构如下：第 2 节提供了迭代处理、记忆机制、注意力机制和位置编码的背景信息。第 3 节详细阐述了新 RT 方案的设计、原则和流程。第 4 节介绍了其有效性的实验验证和分析。第 5 节讨论了其影响和未来工作，第 6 节总结了本文。

## 2. 背景与相关工作

### 2.1 迭代处理与神经网络中的循环机制

神经网络中的迭代计算允许模型在多个步骤中精炼其输出或内部状态，类似于人类的推理

或问题解决过程。这与固定深度的前馈网络形成对比。传统的循环神经网络（RNNs）通过在序列上重复应用相同的循环单元来固有地执行迭代处理 [1]。这使得模型能够根据序列长度进行自适应计算深度 [1]。

一些 Transformer 变体，如通用 Transformer，通过在多个层中重复应用相同的 Transformer 块来引入“深度方向循环” [1]。这种参数共享使得模型能够灵活适应任务复杂性，并通过迭代前向传播有效增加模型的“深度” [1]。通常采用动态停止机制来根据输入复杂性控制迭代次数 [1]。其他方法，如时间潜在瓶颈（Temporal Latent Bottleneck），则采用“块级时间循环”，迭代处理子序列，并使用内部记忆来整合来自先前块的信息 [1]。

新 RT 方案的核心理念是在相同的输入序列上（附加一个记忆标记）执行多次前向传播以精炼最终嵌入。这与深度方向循环（通用 Transformer）的概念紧密对齐，其中相同的计算块被迭代应用 [1]。虽然新 RT 方案并非在传统意义上跨“层”共享权重，但它重用了整个模型的前向传播进行迭代精炼。这表明新 RT 方案隐式地利用了自适应计算深度的优势，允许模型在不增加物理层数的情况下，对复杂任务进行“更深入的思考”。这种联系为多次前向传播可能带来益处提供了坚实的理论基础。它不仅仅是任意的重复，而是一种计算深度的形式，可以像通用 Transformer 中的动态停止机制一样，通过 iter 参数进行自适应控制。

### 2.2 上下文信息记忆机制

记忆增强神经网络（MANNs）旨在保留和利用过去的信息或外部知识来指导未来的计算，解决了传统 RNN 和 Transformer 在处理长期依赖方面的局限性 [2]。记忆组件可以作为中间层集成到 Transformer 架构中，特别是在多头注意力层面，以调节标记选择并高效处理数据 [3]。例如，Transformer-XL 和 MemFormer 使用记忆来存储过去的信息，以处理更长的上下文 [3]。最近的研究也探索了特殊“记忆标记”的使用，其嵌入被优化以表示或重建特定信息 [4]。这些标记可以作为任意文本序列的密集向量表示，并可被优化以存储和检索序列 [4]。

与那些可能使用大型外部记忆块或段级循环（如 Transformer-XL）来存储大量历史信息 [3] 的传统记忆增强神经网络不同，新 RT 方案在每次迭代中生成一个单一的记忆标记 5。这个标记并非预训练用于存储固定序列 [4]，而是根据当前前向传播的注意力和嵌入*动态生成*的 [5]。这种动态生成的、单标记的“摘要”作为当前输入最显著信息的紧凑、注意力引导表示，仅影响下一次前向传播。这种设计选择高效且有针对性。它避免了大型记忆系统的复杂性和潜在噪声，同时为模型的后续“思考”过程提供了集中的“提示”或“摘要”。这是一种内部的自我蒸馏或自我引导形式，而非外部记忆增强。

### 2.3 信息聚合的注意力机制

注意力机制是现代自然语言处理模型，特别是 Transformer 的核心，它允许模型在计算表示时权衡输入序列不同部分的重要性 [6]。自注意力，或称内部注意力，关联单个序列内的不同位置，以计算该序列的表示 [6]。在许多基于 Transformer 的模型中，特定标记被指定用于聚合整个序列的信息，以完成分类或句子表示等任务。例如，BERT 中的

标记通常用于此目的，其输出表示旨在聚合整个序列的信息 [7]。注意力权重直接反映了模型对不同标记的关注程度，表明它们对给定任务或输出的感知重要性 [8]。选择哪个标记的注意力用于聚合至关重要 [9]（指出 并非总是代码表示的最佳标记）。

新 RT 方案特别使用最后一层 <endoftext> 标记的注意力权重来生成记忆标记 [5]。<endoftext> 标记作为序列终止符，其性质决定了它在单向模型中能够看到并聚合所有先行标记的信息 [5]。因此，它的注意力权重自然地表示了它如何“关注”每个标记以形成最终的序列表示。这为推导反映输入序列中标记整体重要性的“注意力引导的输入序列摘要”提供了一个理论上合理的选择 [5]。这与任意平均或使用可能无法完全看到序列的标记的注意力形成对比。这种设计选择提供了一种鲁棒且语义上有意义的方式来创建记忆标记，确保它真正反映了模型对输入的聚合理解，而非嘈杂或部分的表示。它利用了单向架构中 <endoftext> 标记固有的信息聚合能力。

### 2.4 Transformer 模型中的位置编码

Transformer 模型本质上是置换等变的，这意味着它们本身缺乏关于序列中标记顺序的信息 [10]。位置编码对于赋予 Transformer 这种关键的位置和顺序信息至关重要 [10]。正确的位置编码对于自然语言处理任务中的最先进性能至关重要 [10]。不正确地处理 position_ids 会严重降低性能，尤其是在处理比训练时更长的输入序列时（长度泛化问题） [10]。这是因为许多现有位置编码方案无法泛化或引入偏差 [10]。绝对位置编码（APE）和相对位置编码（RPE）是常见的方法 [10]。RPE，如 T5 的，旨在更好地泛化，但可能存在局限性 [10]。

新 RT 方案明确指出，在迭代过程中，position_ids 保持正常的、严格递增的自然数序列 [5]。研究强调，不正确的 position_ids 处理会导致性能下降和模型表征空间的破坏 [10]。通过确保 position_ids 遵循基座模型预期的逻辑，新 RT 方案防止了嵌入空间的“破坏”，并保持了模型正确解释标记顺序和上下文的能力。这不仅仅是一个技术细节，而是一个基本的设计原则，确保新 RT 方案在基座 Transformer 模型已建立的、鲁棒的表征框架内运行。这是新方案在受控性能改进方面取得成功，而没有出现那些原则性较差的迭代方法中灾难性性能下降的关键原因。它突出了在增强预训练模型时所需的微妙平衡。

## 3. 新精炼思想（RT）方案

### 3.1 核心原则与设计理念

新 RT 方案的根本目标是在测试时计算期间使模型能够执行“进一步思考”，利用先前的正向传播“记忆”来重新审视输入序列并生成更高质量的句子嵌入 [5]。指导原则是设计迭代过程，使其最大限度地与基座 Qwen3-Embedding-8B 模型的自然特性对齐。这包括尊重其单向注意力机制及其 position_ids 逻辑 [5]。该方案中的“记忆”被概念化为输入序列的单一、注意力引导的摘要，而非直接复制或大型外部存储 [5]。该方案旨在最小化超参数，并确保每个设计选择都由清晰的理论依据支持，避免任意复杂性 [5]。

### 3.2 迭代信息聚合机制

**记忆标记的生成：** 记忆标记通过获取模型最后一层输出中所有注意力头的注意力矩阵中 <endoftext> 标记所在行来生成 [5]。然后，该注意力用于加权模型最后一层输出中每个标记的嵌入 [5]。其原理在于，<endoftext> 标记位于输入序列的末尾，由于单向注意力机制，它能够全面地查看所有先行标记。其注意力权重反映了它如何聚合整个序列的信息以形成最终的序列表示 [5]。这使其成为生成“注意力引导的输入序列摘要”的理想来源 [5]。

**记忆标记的策略性放置：** 生成的记忆标记在随后的前向传播中被放置在输入序列中紧邻 <endoftext> 标记之前 [5]。其原理是，将其放置在开头会导致“蝴蝶效应”，由于单向注意力，会影响每个标记的嵌入，这是不希望的，因为目标是影响最终聚合的嵌入，而非单个标记的嵌入 5。将其放置在末尾（在<endoftext> 之后）会阻止 <endoftext> 标记关注它，使其无效 [5]。将其放置在<endoftext> 之前，允许 <endoftext> 标记结合“输入序列信息”和“上一轮输入序列的摘要信息”进行更好的聚合，同时不扰动原始输入标记的嵌入 [5]。

**position_ids 的精确处理：** 对于第一次前向传播，position_ids 遵循标准的自然数序列：[[0, 1, 2,..., seq_len-1]] [5]。对于随后的传播，当记忆标记被添加时，position_ids 序列简单地扩展一个位置，保持严格递增的自然数序列（例如，第二次传播为 [[0, 1, 2,..., seq_len]]） [5]。其原理是，这确保了模型对位置信息的固有理解及其表征空间得以保留 [5]。不正确的 position_ids 会严重破坏模型解释输入的能力并导致性能下降 [10]。

### 3.3 新 RT 方案流程示意

**初始化（iter=1）：** 第一次前向传播与基座 Qwen3-Embedding-8B 模型完全相同，使用标准 position_ids 处理输入序列 [5]。这确保了当不需要“进一步思考”时，该方案默认保持基座模型的性能。

**迭代精炼（iter > 1）：**

1. 在初始前向传播结束后，使用 <endoftext> 标记的注意力和最后一层的标记嵌入生成记忆标记 [5]。
2. 对于下一次迭代，将此记忆标记添加到原始输入序列中，紧邻 <endoftext> 标记之前。

3. position_ids 和 attention_mask 根据新的序列长度进行调整与扩展 [5]。
4. 使用此增强输入执行新的前向传播。
5. 重复此过程（生成新的记忆标记并执行另一次前向传播），直到达到指定的迭代次数 iter [5]。
6. 最终的句子嵌入取自最后一次迭代中 last hidden state 的 <endoftext> 标记的输出。

### 3.4 优点与问题解决

新 RT 方案是确定性的，确保了可复现和可预测的行为 [5]。通过使用<endoftext> 标记的注意力，该方案在单向模型中正确捕获了标记对于整体序列聚合的重要性，克服了任意平均的问题 [5]。单一的、注意力引导的记忆标记，经过策略性放置，有效地指导模型随后的信息聚合，提供有意义的上下文，而非仅仅是噪声扰动 [5]。严格遵守基座模型的 position_ids 逻辑和嵌入空间，确保迭代过程不会破坏模型学习到的表示 [5]。该方案的主要超参数是 iter（前向传播次数），其影响是可预测且有理论支持的，从而简化了调优和解释 [5]。该方案旨在实现真正的"进一步思考"，而非表面上的性能提升，这由其在不同复杂性任务上的一致行为所证明 [5]。

## 4. 实验验证与结果

### 4.1 评估任务与指标

新 RT 方案在 C-MTEB 基准测试中的一组语义文本相似性（STS）任务上进行了评估，这些任务被选择以代表不同级别的语义复杂性和推理要求 [5]。

- **PAWSX：** 一个相对简单的 STS 任务，涉及简短的、释义后的句子，带有二元标签 [5]。此任务适用于评估该方案在要求较低的场景中是否保持性能并避免退化。
- **STSB：** 一个更细致的 STS 任务，针对短文本对提供 6 级相似度评分，需要更精细的语义理解 [5]。此任务有助于评估该方案提供细微改进的能力。
- **STS_Algorithm_jd2cv (JD2CV)：** 一个更复杂的、领域特定任务，侧重于职位描述与候选人简历的匹配，需要更深入的语义和上下文理解 [5]。此任务对于证明该方案在需要显著"进一步思考"的场景中的实用性至关重要。

这些 STS 任务的主要评估指标是 Spearman 秩相关系数（Spearman correcoef），这是衡量基于排名的相似性任务的标准指标 [5]。

### 4.2 预期性能特征

- **基座模型性能 (iter=1)：** 预期当 iter=1 时，新 RT 方案将与基座 Qwen3-Embedding-8B 模型表现相同，作为基线 [5]。
- **简单任务（例如 PAWSX）：** 对于需要最少"思考"或复杂推理的任务，增加 iter 预期只会产生微小变化，可能略有下降，但不会出现显著的性能退化。这表明即使在额外

迭代并非严格必要时，该方案也不会破坏模型的表征空间 [5]。

- **复杂任务（例如 STSB，JD2CV）：** 对于需要更深层次语义理解或推理的任务，预期增加 iter 将导致性能逐步提升，达到最佳点，然后可能在此点之后略有下降。这种趋势反映了迭代精炼的益处，直到达到一个点，在此点之后，在不生成新信息的情况下，额外的传递会带来收益递减 [5]。

### 4.3 实验结果

**表 1：新 RT 方案在 PAWSX 任务上的性能（Spearman 相关系数）**

| 前向传播次数 | PAWSX |
|---|---|
| 1（即 Qwen3-Embedding-8B） | 54.37 |
| 2 | 53.79 |
| 3 | 52.94 |

此表展示了新 RT 方案在简单任务上的行为。随着 iter 的增加，性能略有下降（从 54.37 降至 52.94），这种受控的下降证实了迭代过程不会破坏模型底层的表征空间。这证明了该方案的稳定性及其对基座模型完整性的遵守，即使在"进一步思考"并非明显有益的情况下也是如此。

**表 2：新 RT 方案在 STSB 任务上的性能（Spearman 相关系数）**

| 前向传播次数 | STSB |
|---|---|
| 1（即 Qwen3-Embedding-8B） | 86.27 |
| 2 | 86.30 |
| 3 | 86.36 |
| 4 | 86.48 |
| 5 | 86.57 |
| 6 | 86.59 |
| 7 | 86.49 |

此表极具价值，因为它展示了新 RT 方案在*中等复杂*任务上的有效性。观察到的性能趋势

是先提升（从 86.27 到 iter=6 时的 86.59），随后略有下降（iter=7 时的 86.49），这与“进一步思考”的理论预期完美吻合 [5]。这表明迭代处理确实有助于模型精炼其理解，直至达到最佳点，在此之后，在没有新外部信息的情况下，额外的传递会带来收益递减。这验证了该方案的核心假设。

**表 3：新 RT 方案在 STS_Algorithm_jd2cv 任务上的性能（Spearman 相关系数）**

| 前向传播次数 | STS_Algorithm_jd2cv |
| --- | --- |
| 1（即 Qwen3-Embedding-8B） | 62.75 |
| 2 | 65.86 |
| 3 | 68.54 |
| 4 | 70.56 |
| 5 | 72.59 |
| 6 | 73.49 |
| 7 | 74.86 |
| 8 | 75.9 |
| 9 | 75.9 |
| 10 | 75.09 |
| 11 | 74.85 |
| 12 | 74.59 |

此表可能是新 RT 方案在高度复杂任务上有效性的最有说服力的证据。显著且持续的性能提升（从 62.75 到 iter=8, 9 时的 75.9）清楚地表明，迭代精炼显著受益于需要深入理解和复杂信息聚合的任务，例如人岗匹配 [5]。在 iter=10 之后的缓慢下降进一步强化了“最佳点”的特性，表明模型通过重复传递已从输入中提取了大部分可用信息。这为该方案真正增强复杂应用模型性能的能力提供了强有力的经验证据。

### 4.4 有效性分析

新 RT 方案在 PAWSX、STSB 和 JD2CV 任务上的实验结果一致验证了其预测的性能特征 [5]。PAWSX 上最小的性能变化证实了新 RT 方案不会对模型的表征空间引入有害噪声或扭

曲。STSB 和特别是 JD2CV 上显著的性能提升表明，迭代过程成功地使模型能够执行"进一步思考"，从而为需要更深层次语义分析的任务生成更精细、更准确的句子嵌入 [5]。观察到的性能在达到最佳迭代次数后趋于平稳并略有下降，这表明该方案提供了一个受控的精炼过程。模型有效地从输入中提取和聚合信息，在不生成新内容的情况下达到饱和点 [5]。这突出了该方案在利用现有输入信息方面的效率。总而言之，这些结果提供了强有力的经验证据，表明新 RT 方案通过促进有原则的迭代精炼过程，有效地提高了句子嵌入质量，这对于复杂的语义任务尤其有益。

# 5. 讨论与未来工作

## 5.1 贡献总结与有效性证明

本文介绍了一种新颖且有理论基础的迭代精炼句子嵌入方法——新 RT 方案。通过在不同复杂性任务上的严格实验验证，证明了其有效性，在复杂任务上显示出持续的性能提升，同时在简单任务上保持了稳定性。该方案对基座模型特性的遵循，特别是对记忆标记生成和 position_ids 的精心处理，已被证明对其成功以及避免先前迭代方法的缺陷至关重要。

## 5.2 更广泛的影响

新 RT 方案的成功为在推理时增强预训练语言模型的能力提供了一个有前景的方向，而无需进行大规模的再训练或架构修改。这种"测试时计算"方法对于计算资源有限或需要自适应推理深度的应用尤其有价值。它进一步强化了这样一个理念：即使是对前向传播进行细微但有原则的修改，也能通过使模型能够更深入地"思考"其输入，从而解锁显著的性能提升。

## 5.3 当前局限与未来方向

**记忆标记嵌入空间的一致性：** 尽管记忆标记的生成是注意力引导的，但其嵌入可能无法完美符合基座模型固有的嵌入空间 [5]。当前记忆标记是通过加权现有标记嵌入和注意力生成的 [5]。虽然这是一种逻辑聚合，但并未明确说明所得向量是否完美对齐于 Qwen3-Embedding-8B 模型预训练时所使用的单个原始标记嵌入的语义空间。如果存在细微的不匹配，即使整体有益，也可能引入轻微的扰动。这是在预训练模型中注入新信息或修改表示时常见的挑战。未来的工作可以探索确保记忆标记嵌入明确映射或正则化，以更好地符合基座模型原始嵌入空间的方法，可能通过一个小型可训练的投影层或自监督目标来实现。这可能进一步增强稳定性和性能。

**迭代参数（iter）的自适应确定：** 最佳迭代次数 iter 因任务甚至特定输入而异，使其成为一个目前需要通过经验确定的超参数 [5]。数据清楚地表明，最佳 iter 值是任务相关的（例如，PAWSX 需要较少迭代，JD2CV 需要更多） [5]。这意味着固定的 iter 值对于多样化的输入或任务而言可能不是最优的。挑战在于动态确定模型何时对给定输入进行了充分的

“思考”，类似于通用 Transformer 中的动态停止机制 [1]。简单地使用信息熵或余弦相似度等指标可能不足，因为输入的“难度”可能取决于任务 [5]。未来的研究应侧重于开发

iter 的自适应停止标准。这可能包括：

- **任务特定验证集：** 从与测试集分布相似的验证集中采样（例如，可以使用验证集），以帮助确定给定任务的合适 iter 超参数 [5]。
- **基于置信度的停止：** 开发允许模型自主决定何时停止对特定输入进行迭代的指标（例如，输出嵌入的稳定性、内部状态的变化或预测置信度分数）。
- **迭代次数的元学习：**训练一个小型元模型，根据初始输入特征或早期迭代输出预测最佳 iter 值。

## 6. 结论

本文提出了新精炼思想（RT）方案，一种通过迭代信息聚合增强句子嵌入的鲁棒且理论上合理的方法。通过精心设计记忆标记的生成、其策略性放置以及维护位置编码的完整性，新 RT 方案成功地使预训练语言模型能够进行“进一步思考”。我们的实证结果明确证明了其在复杂语义任务上显著提升性能的能力，同时确保了稳定性和保留了基座模型的表示能力。新 RT 方案代表了开发更智能、更自适应的文本理解系统方面的宝贵进展，为更细致和有效的自然语言处理应用铺平了道路。

### 参考文献


1. Investigating Recurrent Transformers with Dynamic Halt - arXiv, 访问时间为 七月 15, 2025， https://arxiv.org/pdf/2402.00976
2. [1909.08314] Memory-Augmented Neural Networks for Machine Translation - arXiv, 访问时间为 七月 15, 2025， https://arxiv.org/abs/1909.08314
3. Combining Transformers with Natural Language Explanations - arXiv, 访问时间为 七月 15, 2025， https://arxiv.org/pdf/2110.00125
4. Memory Tokens: Large Language Models Can Generate Reversible Sentence Embeddings, 访问时间为 七月 15, 2025， https://arxiv.org/html/2506.15001v1
5. RT 方案详细分析及改进.docx
6. Attention is All you Need - NIPS, 访问时间为 七月 15, 2025， https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf
7. Text Summarization with Pretrained Encoders - ACL Anthology, 访问时间为 七月 15, 2025， https://aclanthology.org/D19-1387.pdf
8. arXiv:2406.02536v2 [cs.CL] 15 Oct 2024, 访问时间为 七月 15, 2025， https://arxiv.org/pdf/2406.02536?
9. An Exploratory Study on Code Attention in BERT - arXiv, 访问时间为 七月 15, 2025， https://arxiv.org/pdf/2204.10200
10. Functional Interpolation for Relative Positions Improves Long ..., 访问时间为 七月

15, 2025， https://arxiv.org/pdf/2310.04418
11. Learning interpretable positional encodings in transformers depends on initialization - arXiv, 访问时间为 七月 15, 2025， https://arxiv.org/html/2406.08272