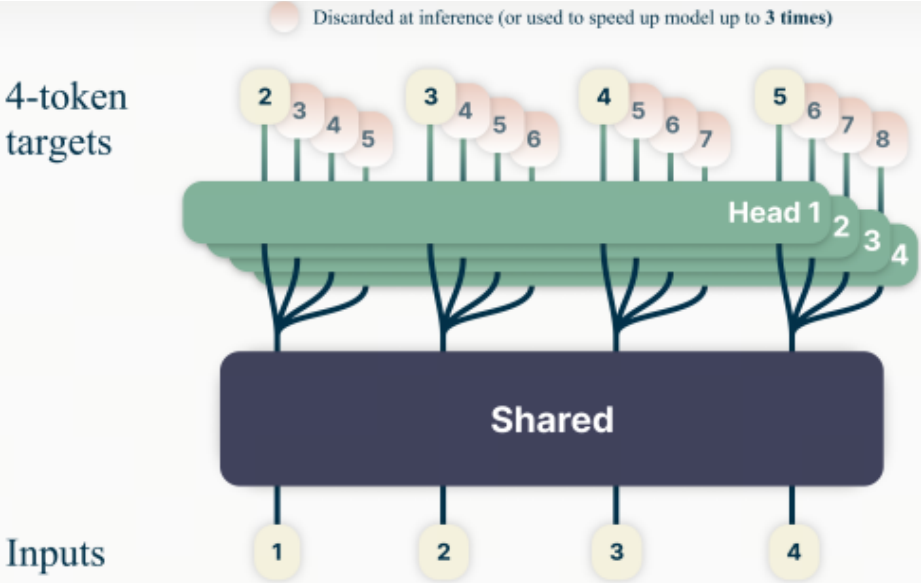
知乎



Meta 2024：如何利用多Token技术，让LLM模型飞速跃升至新境界

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我

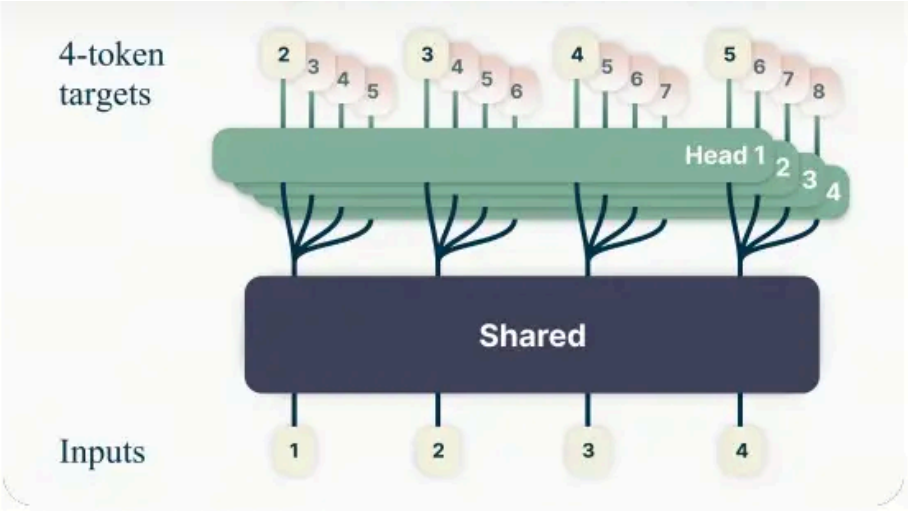
已关注

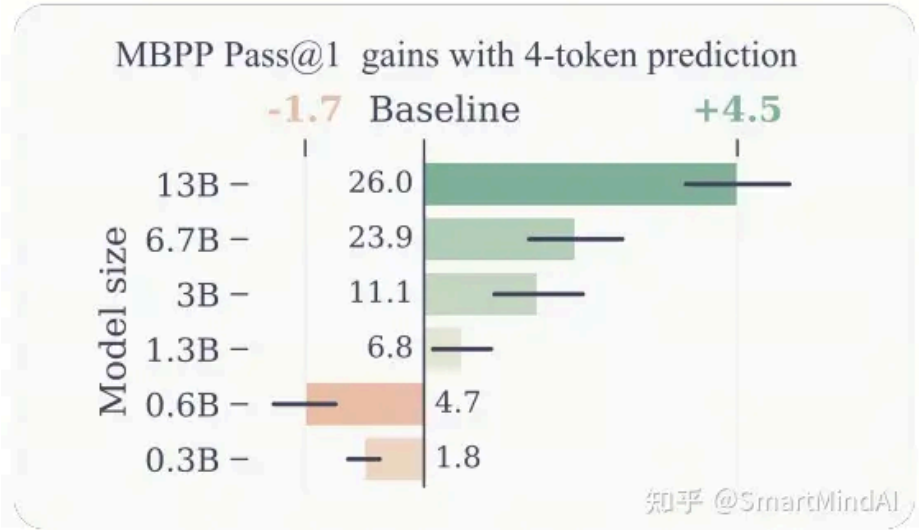
17 人赞同了该文章

Introduction

人类将其最杰出的探索、惊人发现和精美作品凝聚于文字之中。大型语言模型（LLMs），经过对这些大量文本数据的学习，能够提取出惊人的世界知识，并通过执行一项简单却强大的无监督学习任务------即下一次令牌预测，实现了基本的推理能力。然而，这种方法存在局限，依赖"教师强迫"，过度聚焦于局部模式，而非复杂的决策过程。因此，事实仍然是，最先进的下一次令牌预测器往往需要比人类儿童多出几个数量级的数据才能达到同等的流利程度。

知乎





本研究提出，通过引导大型语言模型同时预测多个后续令牌，以提高样本效率。

1. 本研究提出了一种创新的多令牌预测架构，强调了其在不增加额外训练时间和内存消耗的情况下提升模型效率的优点。这种方法通过在现有模型基础上扩展输出头，让模型一次预测多个未来令牌，以此增强模型的理解和泛化能力。
2. 我们通过大规模实验证明，新提出的训练范式在拥有130亿参数的大型模型上确实带来了积极效果。结果显示，这种模型在解决代码问题上的成功率提高了大约15%，这表明它在优化代码理解方面具有显著优势。这种增益并不伴随额外的训练时间和内存消耗，使得这种方法在实践中更具吸引力。
3. 多令牌预测促进了模型的自我推测性解码，这导致在不同批量处理情况下，模型的推理速度有了显著提升，平均快了3倍。这种技术不仅提高了效率，而且在保持或增强模型性能的同时，减少了资源消耗，对实时或大规模应用非常有益。

Method

标准的语言建模通过训练模型预测下一个令牌（x_{t+1}），通常通过最小化交叉熵*损失来学习大型文本序列$(x_1, \ldots, x_T)$。

$$L_1 = -\sum_t \log P_\theta(x_{t+1} \mid x_{t:1}),$$

使用数学表达式，语言建模的目标函数*定义为：

$$L(\theta) = -\sum_{t=n}^{T} \log P_\theta(x_{t+1:t+n} | x_{t-n:t})$$

其中：-θ表示模型参数，-T是序列的总长度，-n是预测的令牌数，-$x_{t-1:t}$是前t个令牌的历史输入，-$x_{t+1:t+n}$是目标连续的n个后续令牌，-$\log P_\theta$是模型对这些目标序列概率的估计。 这个

知乎

这是标准的无监督学习任务，通过训练来优化模型参数θ，使其更好地理解和生成自然语言序列。

$$L_n = -\sum_t \log P_\theta(x_{t+n:t+1} \mid x_{t:1})$$

$$L(\theta) = -\sum_{t=1}^{T-n+1} \sum_{i=1}^{n} \log P_\theta^{(i)}(x_{t+i}|z_{t:1})$$

其中$P_\theta^{(i)}$代表第i个预测头，它依据共享底层的$z_{t:1}$独立推断每个未来令牌x_{t+i}。这种设计允许模型同时处理多个预测，从而既提高了样本利用效率，又确保了预测的精确性。

$z_{t:1} = f_s(x_{t:1})$：由主干生成潜在表示

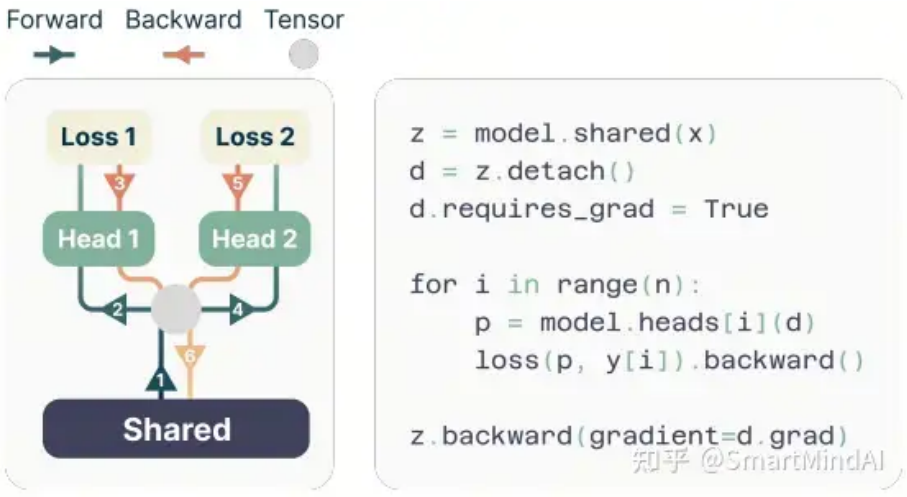
$\hat{x}_{t+i} = f_u(z_{t:1})$：通过未嵌入映射*对每个未来位置进行估计。

$P_\theta^{(i)}(x_{t+i}|z_{t:1}) = softmax(f_{h_i}(\hat{x}_{t+i}))$：利用输出头，以softmax函数的形式预测每个令牌的概率分布。 d 这样的设计使得模型能同时处理多个预测，显著提高了预测速度和精度。

$$P_\theta(x_{t+i} \mid x_{t:1}) = \text{softmax}(f_u(f_{h_i}(f_s(x_{t:1}))))$$

在实践中，针对每个预测单元（$i = 1, 2, \ldots, n$），我们采用如下方法。特别地 $P_\theta(x_{t+1}|x_{t:1})$

是单个下一个令牌预测头。关于多令牌预测的优化，我们注意到GPU内存管理是个挑战，因为词汇表*大小远大于隐状态维度。为节省空间，我们设计了内存高效的方法。在前向传播中，我们逐个处理每个输出头f_i，在主干上累积梯度，而不将整个(n, V)形状的logits和梯度存储。这将内存使用从$O(nV + d)$减少到$O(V + d)$，同时保持运行时间不变（见表格）。



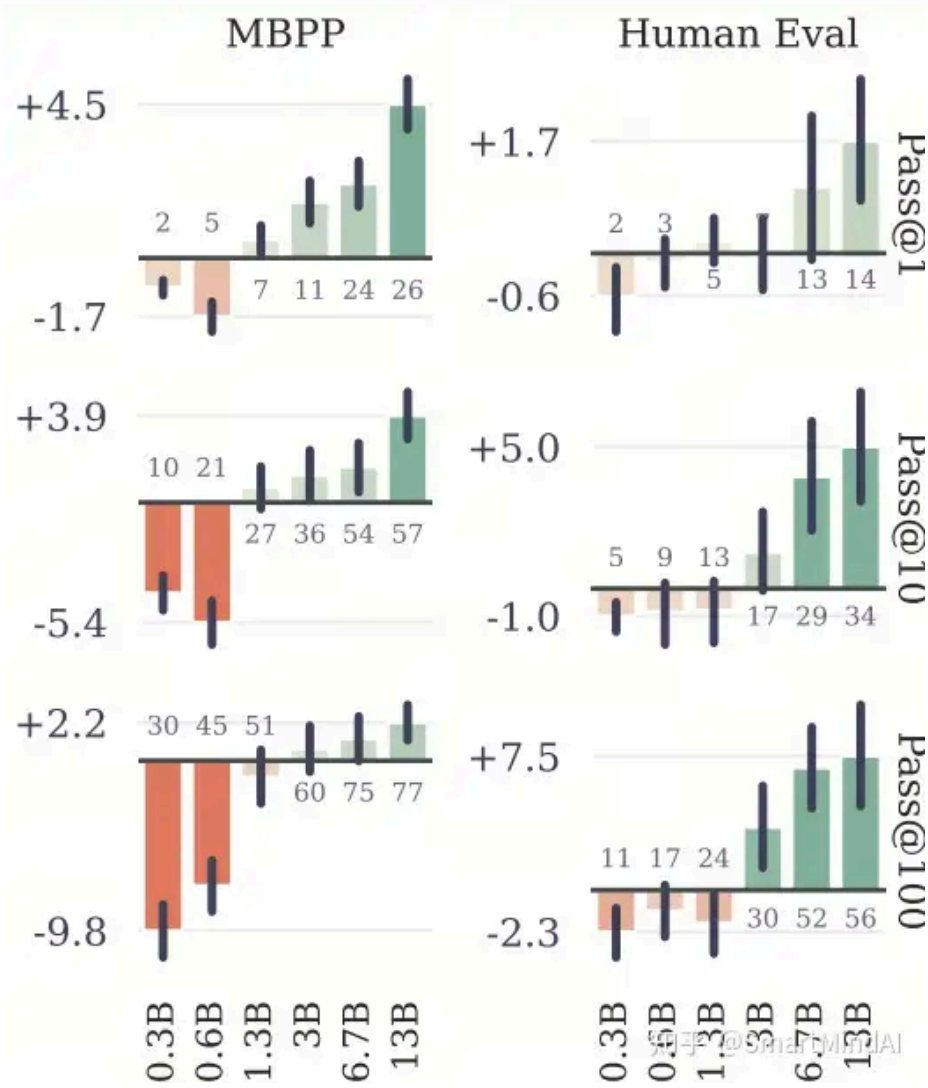
在推理阶段，我们通常用作自回归预测，通过 $P_\theta(x_{t+1}|x_{t:1})$

额外的输出头可以通过块级并行解码或Medusa式树注意力等技术加速解码过程。

Experiments on real data

我们通过七项大规模实验验证了多令牌预测的优势。首先，随着模型规模增大，多令牌预测的有效性愈发凸显。其次，额外的预测头通过推测解码显著提升了推理速度，每轮快达3倍。实验还证实，多令牌预测尤其在字节级分词任务中能更好地捕捉长程依赖。具体到32k大小的分词器，4个预测头带来显著性能提升。在多轮训练中，这种优势持续存在。通过在CodeContests数据集上微调，我们展示了多令牌预训练损失能生成丰富表示。最后，尽管在生成任务如摘要生成中表现出积极影响，但多令牌预测对多项选择题和标准基准的负对数似然并未造成显著退步，保证了性能的平衡。

Benefits scale with model size



图中，MBPP与HumanEval的评估结果展示了在相同计算资源下，相较于单令牌预测，多令牌预测在大型模型上能带来显著性能提升，这体现了在大规模数据下的有效性。我们推测，这种"仅在大规模下有效"的特性可能是为何多令牌预测长期以来未能广泛应用，作为大型语言模型训练的主要损失函数的原因。

Faster inference

我们运用贪婪的自我推测解码，结合xFormers技术和异构批量大小，对70亿参数的4个令牌预测模型进行了实验。在处理测试集代码和自然语言提示的任务中，结果显示，解码速度分别提升3.0倍和2.7倍。对于8字节的预测，推理速度增长了6.4倍（见表）。多令牌预训练使得附加的头部比单独微调下一个令牌模型更精确，从而充分利用了自我推测解码的优势，展示了其显著的准确性提升潜力。

Learning global patterns with multi-byte prediction

我们通过字级别分词的严苛环境测试，证明了多令牌预测确实聚焦于局部模式。使用70亿参数的字级Transformer，处理海量116亿字节（相当于314亿个字符）的数据，发现8字节预测模型在与单一下一个字预测对比下，效果显著提升。具体数据显示，MBPP pass\@1的问题解决率提高了67%，HumanEval pass\@1的问题解决率增长了20%。 这些结果强调，多字节预测不仅是提高训练效率的有效方法，还能显著减少处理长序列的计算成本。相较于仅预测下一个令牌，8字节预测模型通过自我推测解码实现了6倍的运行速度提升，而且在性能上并不逊色于基于令牌的模型，即使使用的是数据量的一半。

Searching for the optimal n

知乎

Training data	Vocabulary	n	MBPP			HumanEval			APPS/Intro		
			@1	@10	@100	@1	@10	@100	@1	@10	@100
313B bytes (0.5 epochs)	bytes	1	19.3	42.4	64.7	18.1	28.2	47.8	0.1	0.5	2.4
		8	**32.3**	**50.0**	**69.6**	**21.8**	**34.1**	**57.9**	**1.2**	**5.7**	**14.0**
		16	28.6	47.1	68.0	20.4	32.7	54.3	1.0	5.0	12.9
		32	23.0	40.7	60.3	17.2	30.2	49.7	0.6	2.8	8.8
200B tokens (0.8 epochs)	32k tokens	1	30.0	53.8	73.7	22.8	36.4	62.0	2.8	7.8	17.4
		2	30.3	55.1	76.2	22.2	38.5	62.6	2.1	9.0	21.7
		4	**33.8**	**55.9**	**76.9**	**24.0**	**40.1**	**66.1**	1.6	7.1	19.9
		6	31.9	53.9	73.1	20.6	38.4	63.9	**3.5**	**10.8**	**22.7**
		8	30.7	52.2	73.4	20.0	36.6	59.6	3.5	10.4	22.1
1T tokens (4 epochs)	32k tokens	1	40.7	65.4	83.4	31.7	57.6	83.2	**5.4**	**17.8**	**34.1**
		4	**43.1**	65.9	83.7	31.6	57.3	**86.2**	4.3	15.6	33.7

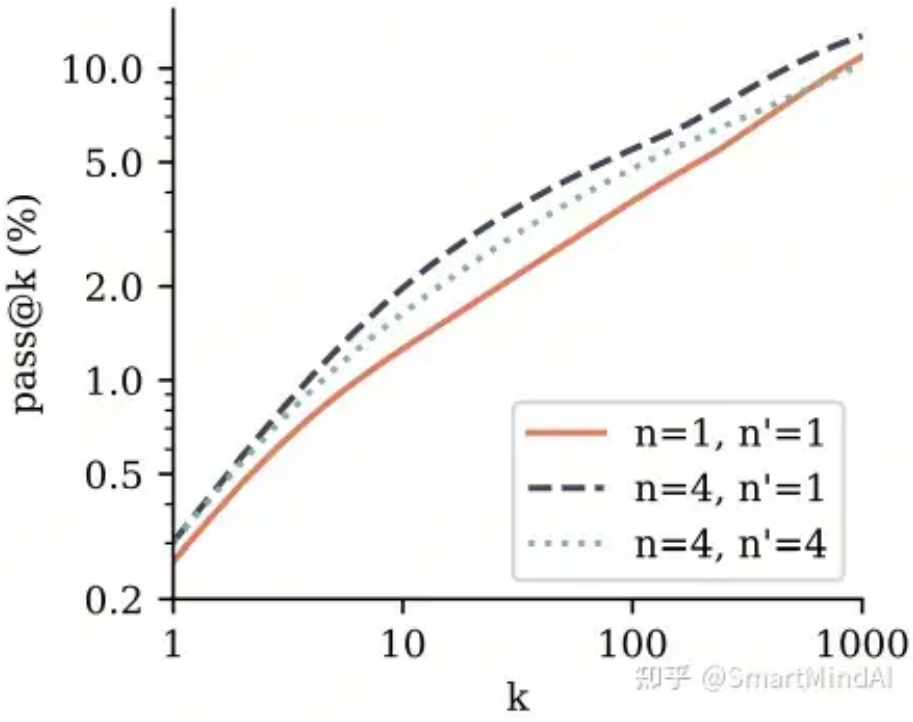
结果显示，不论在MBPP和HumanEval的1、10、100评估指标下，预测4个未来令牌的模型始终表现出最优表现，超越了其他所有设置。

Training for multiple epochs

在相同的训练数据上，多轮训练中，多令牌训练的优势依旧显现，尽管进步幅度有所减小。具体来说，我们在MBPP的pass\@1指标上增益了2.4%，而在HumanEval的pass\@100上增长了3.2%。然而，在某些任务如APPS/Intro中，即使是4个令牌的窗口大小，尽管在2000亿次迭代后，它并非最优化选择，表明仍有改进空间。

Finetuning multi-token predictors

在微调阶段，我们证实了多令牌预测损失预训练的模型在CodeContests数据集上表现出优势，超过了单一的下一个令牌模型。实验对比了4个令牌预测模型（包括一个无额外预测头的版本）、下一个令牌基线，以及两者结合的方案。

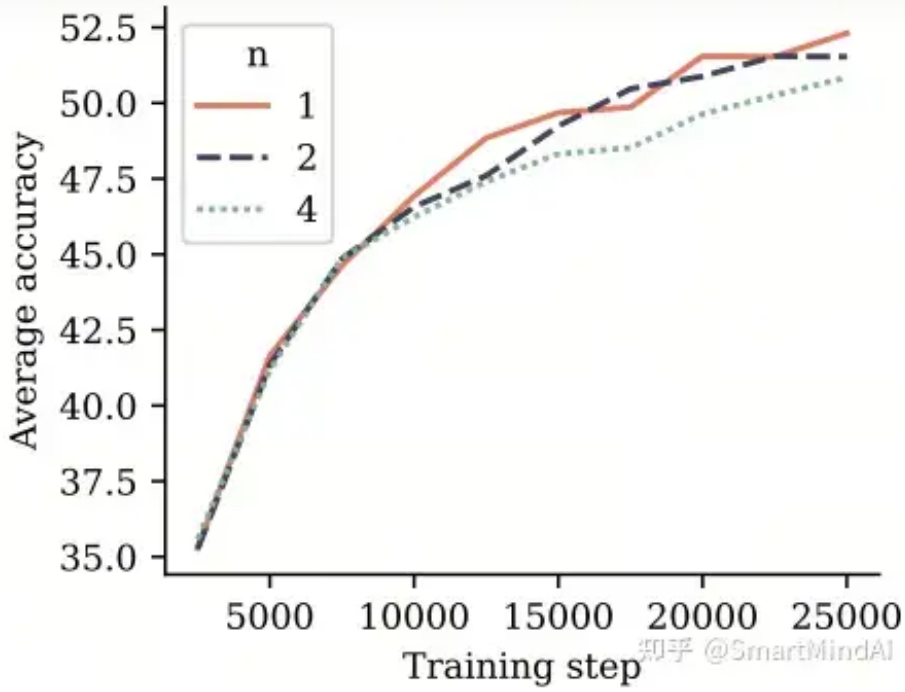


结果显示，不论k值如何，这两种基于4个令牌的预训练方式都明显优于后者，表明它们在理解、问题解决和生成多样答案上有所增强。值得注意的是，CodeContests是最具挑战性的编码基准。

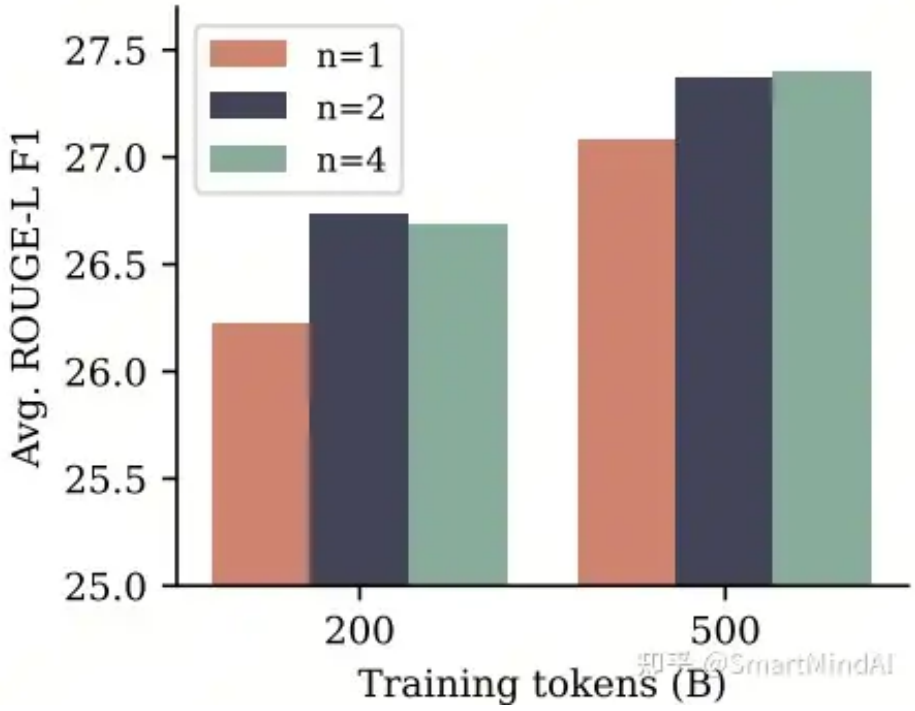
Multi-token prediction on natural language

在探究自然语言中的多令牌预测训练效果时，我们使用了2000亿个句子，共70亿参数的模型，分别训练了4个、2个和下一个令牌损失的模型。图面上，2个令牌预测模型在6个NLP标准基准上的

知乎



然而，我们质疑多项选择和基于概率的评估方法对语言生成能力的准确性。鉴于Koo等人对此类人工标注的局限性提出警示，我们转而通过总结评价和自然语言数学任务来评估模型⁺，避免依赖主观判断。我们比较了2000亿和5000亿步的预训练模型，分别采用下一个令牌和多令牌预测损失，不依赖人类标注。



对于总结评估，我们利用8个基于ROUGE的自动评分基准，对每个模型在每个基准的训练集上微调，每3个周期选取验证集上的最佳ROUGE-L F1分数点。结果显示，不论数据集大小，无论是2个还是4个未来令牌的多令牌预测模型，它们在ROUGE-L F1分数上均优于单一令牌基线，随着数据量增加，两者间的差距逐渐减小。

在自然语言数学的小样本场景（GSM8K）中，我们以8个样本的少样本模式进行评估，关注模型在有限指导下的连贯思考生成答案的精准度。我们通过pass\@k指标来衡量答案的多样性和准确性，测试范围为0.2至1.4的温度。相关结果在附录图表中展示。对于2000亿步的预训练，2个未来令牌模型领先于下一个令牌基线，但在5000亿步和4个未来令牌情况下，这一优势被4个令牌模型所超越，且后者在所有测试中表现都不佳。

知乎

我们提出多令牌预测来解决仅基于下一个令牌训练的问题，特别是在大型模型如代码任务上效果显著。实验表明，结合推测性解码，预测速度可提升3倍。未来，我们计划探索如何自动选择多令牌预测中的n，可能通过损失缩放和平衡策略。我们也将研究适应性地调整多令牌与下一个令牌的词汇大小，以优化计算成本与序列长度的关系。此外，我们还将研发在嵌入空间中的改进辅助预测损失。

原文《Better & Faster Large Language Models via Multi-token Prediction》

发布于 2024-05-07 10:46 · IP 属地北京

LLM Meta分析 meta-learning