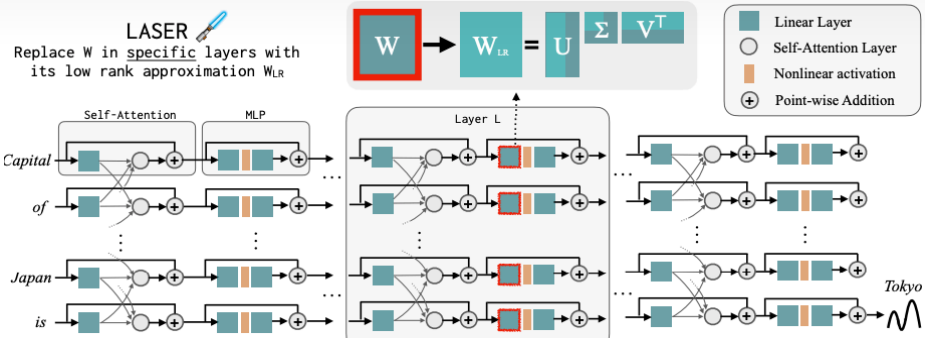
知乎



麻省理工-2023：通过层选择性降级加速LLM大模型的推理性能

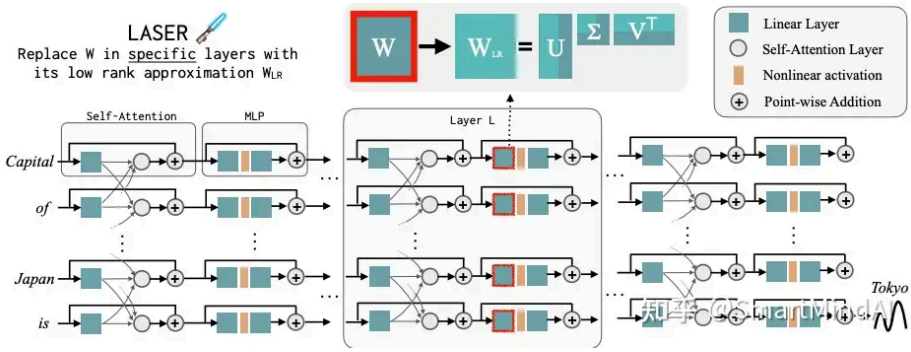
SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

Introduction

Transformer模型在多个关键任务中表现优秀。其底层结构是处理和理解自然语言的最佳选择，并在CV和RL领域具有潜力。目前流行的Transformer版本需要大量计算资源进行训练和推断。这刻意设计，因为训练更多参数或数据的模型通常优于较薄的前代，且优势显著。

然而，越来越多的研究表明，尽管模型有许多可适配参数，但在推断时可以显著修剪，而不会严重影响性能。这一现象提高了泛化与过拟合间关系的关注度，并推动了开发适用于高效推断的剪枝策略的研究。该研究探讨了模型训练数据与**LASER**样本间的关系。发现**LASER**提供了一种降噪过程，有助于弱学习事实的可访问性，并提高对过去正确问题的同义替换的鲁棒性。

本文研究了如何通过删除高层组件来提高模型性能。当只在**LASER**后正确回答问题时，原始模型的回答通常包含高频词汇，这些词汇并非与正确答案完全相同。但在进行了排名降低后，模型的回答就会变为正确。作者将剩余组件单独表示，并使用仅包含其高阶奇异值的权重矩阵进行近似。结果显示，这些组件描述了与正确答案具有相同语义类型的其他响应，或者通用高频词汇。



Related work

本文是首次研究如何通过降低变换器的层次来提高其性能。然而，仍有一些未解决的问题，如如何存储LLM的事实和如何优化网络结构。已有研究表明，模型在检查实体特性时会在不同层中存储事实信息，并且通过微调选择的层可以增强模型对分布偏移的鲁棒性。

然而，对于大型语言模型如何组织和使用这些信息以生成答案，目前存在相互矛盾的证据。一些理论提出信息在变换器模型的全连接网络（MLP）部分以两层键值记忆的形式存储，并通过自注意力模块逐层复制。此外，还有一种方法叫做追踪和编辑特定实体的本地信息，并将其映射到不同的"不可能"输出，这支持了局部理论。同时，早退现象也支持了这种理论，即在中间层的表示可以直接用于模型终端头生成输出。

反例是，Hase（2023）发现模型架构中的某些实体或实体关系信息可以被不同的层修改，因此事实可以在多层级上以分散的形式存储。模型压缩：神经网络剪枝可显著减小存储需求和推理时间。已经发现卷积、全连接和Transformer模型存在稀疏子网络。尽管模型剪枝可以改进模型泛化

知乎

本文发现，选择性剪枝能单独改善模型泛化，而不需要额外的训练。此外，我们发现早期层剪枝可能会导致性能损失，而后期层剪枝则通常会有明显的性能提升。该部分研究了具有三个参数α, β和 γ的复杂非线性系统⁺。这些参数可由实验数据确定，并通过贝叶斯方法更新其不确定性，从而有效地控制系统的动态行为。

Preliminaries

核心组件与基本符号概述 该部分包含我们的研究的核心组件和基本符号的介绍 我们使用$\RR$表示实数，$\NN$表示自然数，小写字母如$v \in \RR^d$表示一个维数为d的向量，大写字母如$W \in \RR^{m \times n}$表示大小为$m \times n$的矩阵。我们使用$\|v\|_2$来表示向量v的欧几里得⁺范数，$\|W\|_2$来表示矩阵W的谱范数⁺。我们使用$[N]$表示集合$\{1, 2, \cdots, N\}$。

我们将使用$\mathbf{rank}(W)$来表示矩阵W的秩，并且将使用$\sigma_i^{\downarrow}(W)$来表示其第i个最大的奇异值。提供了一种简洁的Transformer架构描述，用于我们的分析。该架构由L层Transformer块组成，每层包含自注意力机制和前馈网络，用于混合时间步信息并处理每个时间步信息。我们的目标是提供基本术语，而不是详细调查不同模型之间的实现差异。

自注意力机制通过一个查询向量q和三个相关向量h, k, v计算出注意力概率p(i, j)，用于计算注意力向量z(i)。注意力向量z(i)是根据所有元素的概率计算得到的。自注意力机制再通过投影矩阵⁺o(w)把注意力向量z(i)和h(i)相加得到最终结果u(i)。前馈步骤通过两层具有ReLU或GELU激活函数的多层感知器（MLP），将每个向量应用到每个输入令牌上。通常，这些MLP的权重矩阵分别为U_{in}和U_{out}。每一层变换器块的输出是通过$\psi(\cdot)$和原始输入的和得到的。Transformer架构的权重矩阵包括查询权重矩阵W_q，键权重矩阵W_k，值权重矩阵W_v，输出权重矩阵W_o，嵌入输入令牌的嵌入矩阵，以及最后一层之前的投影权重矩阵。

我们将在研究中主要关注这些权重矩阵，并对其进行干预。矩阵$W \in \RR^{m \times n}$和$r \in \NN$的问题要求找到一个矩阵$\hat{W}$，使得$\|\cdot\|_2$最小，并且满足

$\mathbf{rank}(\hat{W}) \leq r$。

Eckart-Young-Mirsky定理给出了求解此问题的最优解------使用奇异值分解⁺(SVD)。这个分解可以表示为$W = U\Sigma V^T$，其中

$U = [u_1, u_2, \cdots, u_m] \in \RR^{m \times m}$，

$V = [v_1, v_2, \cdots, v_n] \in \RR^{n \times n}$，

$\Sigma \in \RR^{m \times n}$是一个对角矩阵，其对角元素由$W$的奇异值按降序排列得到。我们也可以将$W$的SVD表示为

$W = \sum_{i=1}^{\min\{m,n\}} \sigma_i^{\downarrow}(W) u_i v_i^T$。

根据Eckart-Young-Mirsky定理，向量组

$\hat{u}_i = \sum_{j=1}^{r} \sigma_i^{\downarrow}(W) u_j v_j^T$

构成的矩阵

$\hat{W} = \sum_{i=1}^{r} \sigma_i^{\downarrow}(W) \hat{u}_i \hat{v}_i^T$

是对给定期望秩$r \leq \min\{m, n\}$的最优解决方案。在这篇文章中，我们将使用"更高阶成分"来指代SVD中的小奇异值对应的项。这些成分通过**LASER**被移除。

LALASER是选择性排名降低的简称。

我们定义了**LASER**操作，其包含三个参数：τ, ℓ, ρ。τ决定了干预的对象类型（如MLP或注意力层），ℓ表示干预的层次编号（从0开始），ρ则控制降维的程度（取值范围在$0, 1$之间）。具体而

知乎

Experiments

研究了**LASER**在Transformer各层的后果，包括动机分析及针对CounterFact问题回答数据集的研究。

A Thorough Analysis with GPT-J on the CounterFact Dataset

Dataset		Model Name					
		Roberta		GPT-J		LLama2	
			LASER		LASER		LASER
CounterFact	Acc	17.3	**19.3**	13.1	**24.0**	35.6	**37.6**
	Loss	5.78	**5.43**	5.78	**5.05**	3.61	**3.49**
HotPotQA	Acc	6.1	**6.7**	**19.6**	19.5	16.5	**17.2**
	Loss	10.99	**10.53**	3.40	**3.39**	3.15	**2.97**
FEVER	Acc	50.0	**52.3**	50.2	**56.2**	59.3	**64.5**
	Loss	2.5	**1.76**	**1.24**	1.27	1.02	**0.91**
Bios Gender	Acc	87.5	**93.7**	70.9	**97.5**	75.5	**88.4**
	Loss	**0.87**	1.13	**3.86**	4.20	3.48	**2.93**
Bios Profession	Acc	64.5	**72.5**	75.6	**82.1**	85.0	**86.7**
	Loss	**4.91**	6.44	**4.64**	4.91	4.19	**4.05**
TruthfulQA	Acc	56.2	56.2	54.9	**55.6**	50.5	**56.2**
	Loss	1.60	**1.42**	1.02	**1.01**	**0.95**	1.04
BigBench-Epistemic Reasoning	Acc	37.1	**41.8**	37.1	**38.3**	44.8	**63.4**
	Loss	9.39	**6.80**	0.74	**0.62**	0.78	**0.73**
BigBench-WikidataQA	Acc	28.0	**30.7**	51.8	**65.9**	59.5	**62.0**
	Loss	9.07	**7.69**	3.52	**2.86**	2.40	**2.31**

Table 1: The effect of LASER intervention on eight natural language understanding datasets. We find the best LASER intervention for each model and task using accuracy/0-1 on a validation set and report its performance on a held-out test set. In some of the cases, while the model's accuracy improves, its loss slightly worsens.

我们通过贪婪搜索(,)在多个层上改进模型性能，从大到小选择ℓ和小到大选择。搜索仅在MLP层上进行，以提高速度并寻找最大改进。搜索在验证集上进行，测试集上报告结果。CounterFact使用GPT-J模型基础0-1精度为13.1%。最佳一步**LASER**使模型准确率达到24.0%。同一层**LASER**将top-10精度提高到29.2%，比单层**LASER**提升5.2%准确度。不同(ℓ, ρ)组合搜索结果可在图中查看。

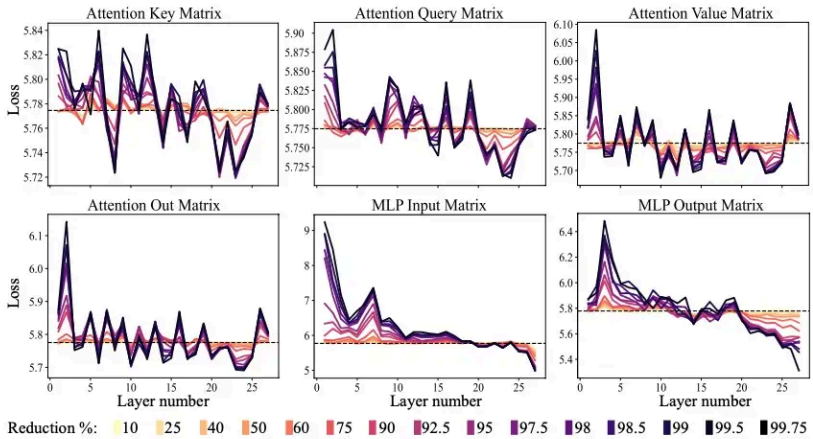


Figure 2: The effect of rank reduction across different layer types is not uniform. Here we show the effect of rank reduction for GPT-J as studied on the CounterFact dataset. The dashed line is the un-modified network's loss. In the attention layers (key, query, value, out matrices), while it is clear matrices could be significantly rank-reduced without damaging the learned hypothesis, there is very little performance increase. However, for the multi-layer perceptron (MLP) layers, rank reduction does not uniformly benefit to improving the model's performance (around layer 20).

Which facts in the dataset are recovered by rank reduction?

为了理解这一现象，我们首先查看了在**LASER**干预后正确回答的问题以及信息与训练数据出现频率的影响。对于CounterFact中的每一个数据点，我们都检索了PILE中包含实体和答案的所有例子。然后，我们计算了每个评估问题在训练数据中出现的次数。 我们发现在排名减少的情况下，最可能很少出现在数据中的是恢复的事实。"原始正确"描述了即使没有任何干预也正确分类的样本。"答案纠正"指的是模型仅在进行**LASER**干预后才能正确的题目。

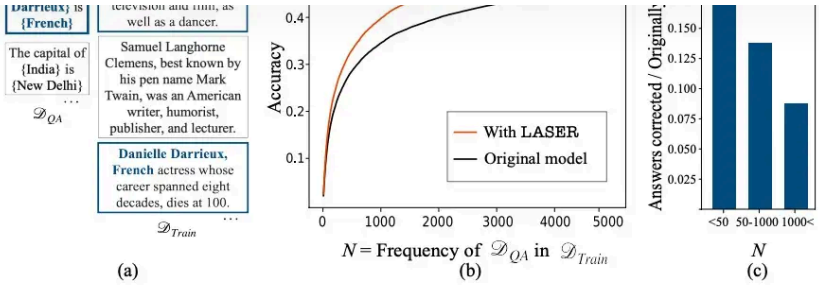


Figure 3: Which datapoints benefit from LASER? We analyze how frequently in the training data "corrected" facts occur. GPT-J is an ideal test bed for such analysis since its training data ($\mathcal{D}_{Train}$), the PILE dataset, is publicly available. (a) For GPT-J evaluated on CounterFact ($\mathcal{D}_{QA}$) we retrieve all the datapoints in $\mathcal{D}_{Train}$ that contain a mention of both the entity of interest and the answer that correspond to each sample in $\mathcal{D}_{QA}$. (b) A plot depicting the cumulative top-10 accuracy of the model on all datapoints that occur in the training data less than or equal to the frequency indicated on the x-axis. Here we show accuracy with and without LASER. (c) The largest boost in performance occurs for low-frequency samples. This bar chart displays the amount of boost offered by LASER for data binned by the frequency with which corresponding facts occur in $\mathcal{D}_{Train}$. Maximal improvements in accuracy are from datapoints that have less-frequent occurrences in training data.

What are higher-order components storing?

本节通过CounterFact数据集和GPT-J研究保留高阶成分是否有助于提高模型性能。研究人员发现，对于问答任务，改善往往出现在答案支持的训练集中出现频率较低的数据的问题上。虽然清楚地消除高阶成分"去噪"了模型并帮助恢复"隐藏"，但不清楚高阶成分代表什么使得删除它们提高了性能。因此，研究人员使用高阶成分近似最终权重矩阵，并分析当执行**LASER**时，模型在GPT-J最初不正确但被翻转为正确的数据点上的行为变化。他们发现在保留前k个组件的情况下，模型对这些问题的答案从通用单词变为正确的实体。同时，随着他们系统地包含低阶成分，模型的输出改变为预测高频单词。研究人员还通过测量平均余弦相似度来调查这种系统的退化，并发现删除高阶成分能够解决内部冲突，使模型准确回答。

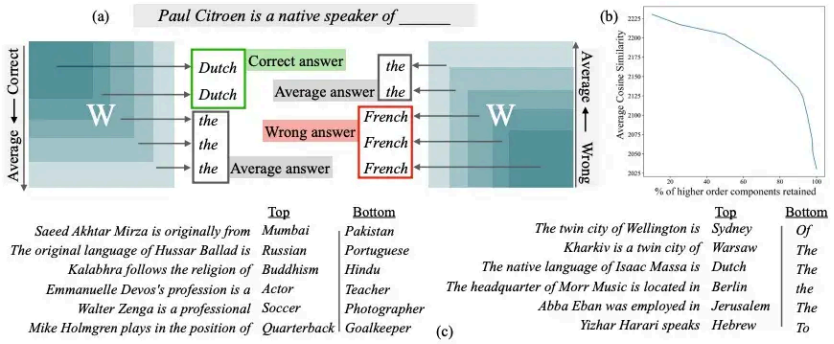


Figure 5: (a) [*Left*] LASER approximates learned matrices by their lower-order components. We find that for datapoints where the model's predictions improve after LASER, if we instead use the entire matrix (including higher-order components), the model often predicts only "generic" words. (a) [*Right*] To understand what these higher-order components encode, we approximate the learned weight matrix with the higher-order components instead. We find that these higher-order components sometimes encode the correct semantic type of the answer but the incorrect response. (b) Analytically, computing the semantic similarity (cosine distance between the true answer and the answers generated by the bottom k% of the singular vectors) shows that on average the answer computed by the higher-order components is more similar to the real answer. (c) Shows some examples from the dataset and the corresponding answers computed by the top-1 and bottom fraction of the components.

How generally does this hold?

我们评估了模型在不同语言理解任务上的表现。我们使用七种数据集：CounterFact、HotPotQA、FEVER、Gender and Profession in Bios、TruthfulQA、BigBench-Epistemic Reasoning和BigBench-WikidataQA。这些数据集评估了语言理解的各种方面。我们还考虑了模型在应对性别和职业偏见方面的性能，通过测试Bias in Bios数据集来衡量偏见。HotPotQA是一个更复杂的开放式问答任务，需要模型具备较长的答案处理能力。BBH的Epistemic Reasoning数据集测试了模型的逻辑推理和阅读理解能力。最后，TruthfulQA测试了模型的准确性。我们使用20%的数据集作为验证集，并选择最优的干预超参数。其余80%的数据集展示了所选超参数的效果。我们使用的模型有罗伯特·奥本海默[+]、GPT-J和LLAMA2，有关数据集的详细信息请参考附录。

(i) 利用LLM生成序列并检查其是否包含目标答案，若无则返回0； (ii) 根据答案在可能值集合中的位置计算分类准确率，若在最可能的答案集中，则判断其更偏向于正确答案； (iii) 对未使用的数据

度下降且可能导致性能提升，但所需降解程度各不相同。

Non-text domains

我们评估了强化学习代理在文本域外使用时的排名下降情况。

Model Name	Acc.	Return
Transformer	50.67	0.575
with LASER	**53**	**0.965**

Table 2: Effect on LASER on a 6-layer Decision Transformer agent. The base model is trained and evaluated in a challenging 10 × 10 Sokoban domain.

我们研究了干预对决策Transformer在Sokoban游戏上的表现的影响，并对其进行了进一步评估。这是一个复杂的规划问题，需要代理将物体移动到洞口上方。决策Transformer的输入为特定状态下的环境视图，输出为基本操作。结果表明，在接受Sokoban训练的决策Transformer模型中，解决的任务数量增加了3%（表）。实验详情可在附录中查看。

Conclusion and Discussion

首先，我们需要理解为什么在训练过程中，更高的阶权重矩阵会产生噪声；其次，我们需要考虑模型架构以及其他因素是否会影响这种现象；最后，我们需要理解为什么这种现象会在MLP的晚期层中特别明显。这些都需要进一步的研究来探索和解答。

原文 《The Truth is in There: Improving Reasoning in Language Models with Layer-Selective Rank Reduction》

发布于 2023-12-28 14:03 · IP 属地北京

LLM　LLM推理　麻省理工