Why Can GPT Learn In-Context?Language Models Implicitly Perform Gradient Descent asMeta-Optimizers

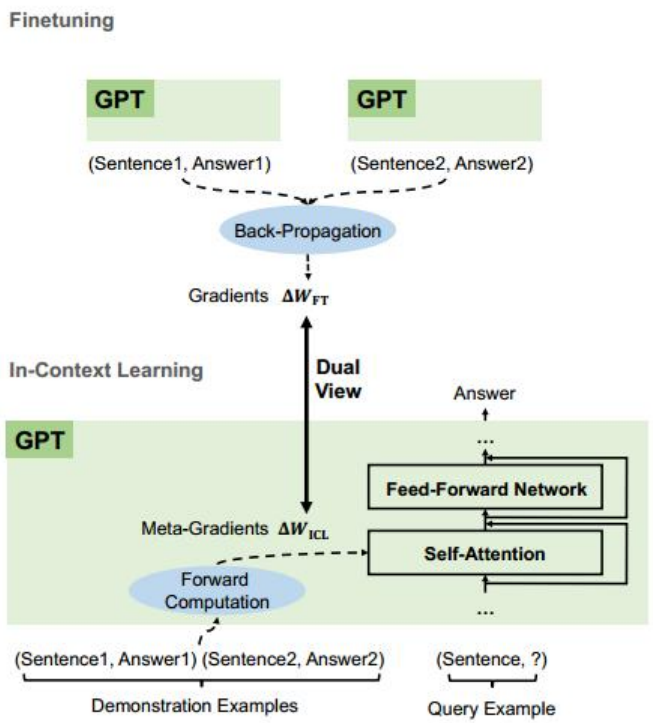
Damai Daiy* , Yutao Sunk* , Li Dongz, Yaru Haoz, Shuming Maz, Zhifang Suiy, Furu Weizy
{daidamai,szf}@pku.edu.cn
{lidong1,fuwei}@microsoft.com

2025.04.17

概述

大型预训练的语言模型显示了令人惊讶的**上下文学习**（In-Context Learning，ICL）能力。通过一些示范性的输入-标签对，它们可以预测未见过的输入的标签，而无需额外的参数更新。尽管在性能上取得了巨大的成功，但ICL的工作机制仍然是一个开放的问题。

为了更好地理解ICL的工作原理，本文将语言模型解释为元优化器，并将ICL理解为一种隐性的微调。从理论上讲，作者分析了Transformer注意力有一个基于梯度下降的优化的双重形式。并在此基础上可以得出如下对ICL的理解：GPT首先根据示范实例产生元梯度，然后将这些元梯度应用于原始的GPT，建立ICL模型。



在本文中，作者将ICL解释为一个元优化的过程，并试图在基于GPT的ICL和微调之间建立联系。作者发现Transformer注意力有一个基于梯度下降的优化的双重形式，并在此基础上提出了一个解释ICL的新视角：

（1）预训练的GPT作为元优化器；

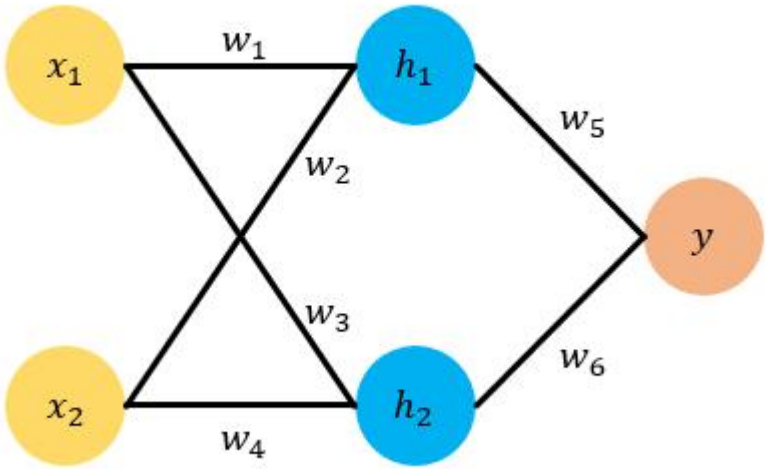
（2）它通过正向计算根据示范例子产生元梯度；

（3）元梯度通过注意力应用于原始语言模型以建立ICL模型。

如图所示，ICL和显式微调共享一个基于梯度下降的优化的双重观点。唯一的区别是，ICL通过正向计算产生元梯度，而微调则通过反向传播计算梯度。因此，将ICL理解为某种隐式微调是合理的。

反向传播

- 反向传播的基本思想就是通过计算输出层与期望值之间的误差来调整网络参数，从而使得误差变小。



- 输入：$x_1 = 1, x_2 = 0.5$
- 目标：$t = 4$
- 权重：$w_1 = 0.5, w_2 = 1.5, w_3 = 2.3, w_4 = 3, w_5 = 1, w_6 = 1$

$$w^+ = w - \eta \cdot \frac{\partial E}{\partial w}$$

$$h_1 = w_1 \cdot x_1 + w_2 \cdot x_2 = 1.25$$
$$h_2 = w_3 \cdot x_1 + w_4 \cdot x_2 = 3.8$$
$$y = w_5 \cdot h_1 + w_6 \cdot h_2 = 5.05$$
$$E = \frac{1}{2}(y - t)^2 = 0.55125$$

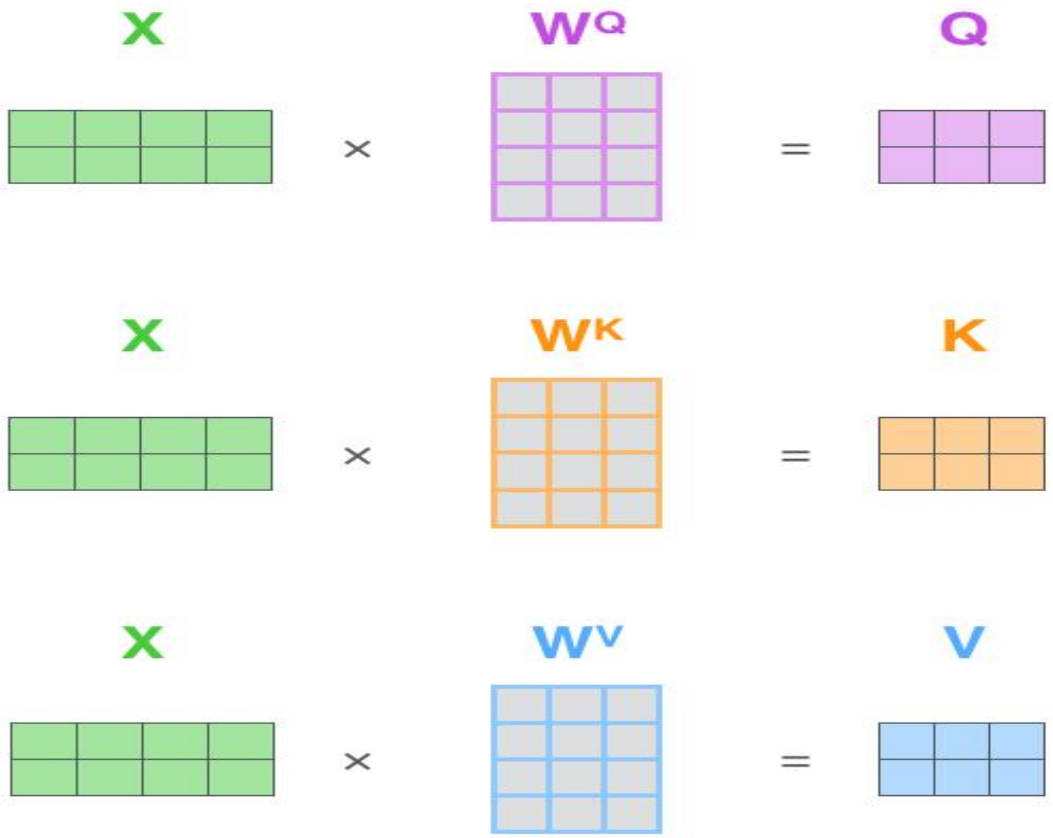
假设 $\eta = 0.1$

$$\frac{\partial E}{\partial w_5} = \frac{\partial E}{\partial y} \cdot \frac{\partial y}{\partial w_5}$$
$$= 2 \cdot \frac{1}{2} \cdot (t - y) \cdot (-1) \cdot h_1$$
$$= 1.05 \times 1.25 = 1.3125$$

$$w_5^+ = w_5 - \eta \frac{\partial E}{\partial w_5}$$
$$= 0.86875$$
$$E^+ = 0.3388$$

注意力机制

- 注意力机制源自人类处理信息的方式。由于外界信息量庞大且复杂，远远超过人脑的处理能力，因此人在处理信息时会优先关注重要的部分，而忽略无关的信息。
- 对于没有情感的机器来说，注意力机制可以理解为一种“赋权”操作。通过为不同信息分配权重，机器能够优先处理更重要的信息，而弱化不相关的内容。



Query（查询）： 表示当前的关注点或目标，是主观意识的体现，自主提示。可以理解为模型在特定时刻的“**需求**”或“**兴趣点**”。

Key（键）： 表示被比对的对象，是客观事物的突出特征，非自主提示。Key是模型用来与Query进行匹配的特征向量，帮助模型决定哪些信息是相关的。Key可以理解为数据的“**索引**”或“**标识符**”。

Value（值）： 表示与Key对应的详细信息，代表物体本身的特征向量。当Query与某个Key匹配时，模型会提取相应的Value作为输出。Value可以理解为数据的“**实际内容**”或“**特征表示**”。

方法

- **基于梯度下降的优化和注意力之间的双重形式**：Transformer在进行in-context learning的前向传播过程中，等价于进行梯度的反向传播更新参数。在一个简化的线性模型中，反向传播梯度更新的过程可以表述为如下式子：

$$\mathcal{F}(\mathbf{x}) = (W_0 + \Delta W)\,\mathbf{x}.$$

$$\Delta W = \sum_i \mathbf{e}_i \otimes \mathbf{x}'_i,$$

其中e表示梯度，x表示输入表示，二者的外积得到参数的梯度，原来的模型参数加上这个梯度变化，得到以此梯度更新后的新参数。

$$\begin{aligned}\mathcal{F}(\mathbf{x}) &= (W_0 + \Delta W)\,\mathbf{x} \\ &= W_0\mathbf{x} + \Delta W\mathbf{x} \\ &= W_0\mathbf{x} + \sum_i \left(\mathbf{e}_i \otimes \mathbf{x}'_i\right)\mathbf{x} \\ &= W_0\mathbf{x} + \sum_i \mathbf{e}_i \left(\mathbf{x}'^T_i \mathbf{x}\right) \\ &= W_0\mathbf{x} + \mathrm{LinearAttn}\left(E, X', \mathbf{x}\right),\end{aligned}$$

第二项就是attention结构，query、key、value分别为当前输入、历史输入和回传的梯度。这个转换表明，我们通过这种数据上的attention进行前向传播，可以达到梯度更新后的参数前向传播相类似的预测结果。

方法

- In-context learning的理解

$$\begin{aligned}\mathcal{F}_{\mathrm{ICL}}(\mathbf{q}) &= \mathrm{Attn}(V, K, \mathbf{q}) \\ &= W_V[X'; X]\,\mathrm{softmax}\left(\frac{(W_K[X'; X])^T \mathbf{q}}{\sqrt{d}}\right)\end{aligned}$$

其中X和X'分别代表原始输入样本和demonstration样本。

$$\begin{aligned}\mathcal{F}_{\mathrm{ICL}}(\mathbf{q}) &\approx W_V[X'; X]\,(W_K[X'; X])^T \mathbf{q} \\ &= W_V X\,(W_K X)^T \mathbf{q} + W_V X'\,(W_K X')^T \mathbf{q} \\ &= \widetilde{\mathcal{F}}_{\mathrm{ICL}}(\mathbf{q}).\end{aligned}$$

省略softmax等非线性操作，上面的公式可以进一步表述成如下形式。

$$\begin{aligned}\widetilde{\mathcal{F}}_{\mathrm{ICL}}(\mathbf{q}) &= W_{\mathrm{ZSL}}\mathbf{q} + W_V X'\,(W_K X')^T \mathbf{q} \\ &= W_{\mathrm{ZSL}}\mathbf{q} + \mathrm{LinearAttn}\,(W_V X', W_K X', \mathbf{q}) \\ &= W_{\mathrm{ZSL}}\mathbf{q} + \sum_i W_V \mathbf{x}'_i \left((W_K \mathbf{x}'_i)^T \mathbf{q}\right) \\ &= W_{\mathrm{ZSL}}\mathbf{q} + \sum_i \left((W_V \mathbf{x}'_i) \otimes (W_K \mathbf{x}'_i)\right) \mathbf{q} \\ &= W_{\mathrm{ZSL}}\mathbf{q} + \Delta W_{\mathrm{ICL}}\mathbf{q} \\ &= (W_{\mathrm{ZSL}} + \Delta W_{\mathrm{ICL}})\,\mathbf{q}.\end{aligned}$$

又可以进一步转换为梯度更新的形式：也就是说，in-context learning的前向传播过程，相当于对原来的ZSL参数进行了梯度更新，而这个梯度就来源于demonstration，通过attention的方式将其应用到当前样本的参数更新上，实现类似finetune的预测效果。

实验

- 作者在基于横跨三个分类任务的六个数据集来比较ICL和微调。SST- 2、SST-5、MR和Subj是四个用于情感分类的数据集；AGNews是一个话题分类数据集；CB用于自然语言推理。

	SST2	SST5	MR	Subj	AGNews	CB
# Validation Examples	872	1101	1066	2000	7600	56
# Label Types	2	5	2	2	4	3
ZSL Accuracy (GPT 1.3B)	70.5	39.3	65.9	72.6	46.3	37.5
FT Accuracy (GPT 1.3B)	73.9	39.5	73.0	77.8	65.3	55.4
ICL Accuracy (GPT 1.3B)	92.7	45.0	89.0	90.0	79.2	57.1
ZSL Accuracy (GPT 2.7B)	71.4	35.9	60.9	75.2	39.8	42.9
FT Accuracy (GPT 2.7B)	76.9	39.1	80.0	86.1	65.7	57.1
ICL Accuracy (GPT 2.7B)	95.0	46.5	91.3	90.3	80.3	55.4

六个分类数据集（第1-2行）的统计数据以及这些数据集上的零样本学习（ZSL）、微调（FT）和上下文学习（ICL）设置的验证准确性。

实验

- ICL涵盖了大多数微调的正确预测

Model	SST2	SST5	MR	Subj	AGNews	CB	Average
GPT 1.3B	91.84	66.67	97.08	87.17	83.08	87.50	85.56
GPT 2.7B	96.83	71.60	95.83	87.63	84.44	100.00	89.39

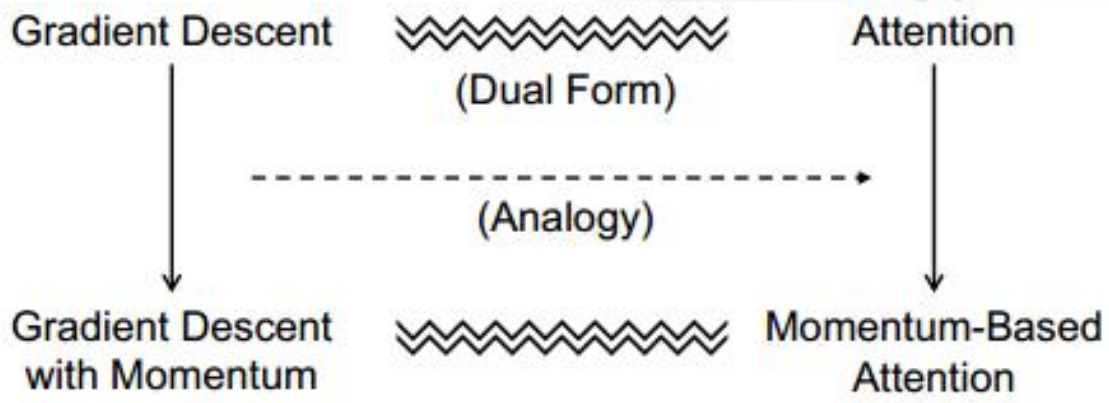
Rec2FTP（微调预测召回率），用于六个数据集上的两个GPT模型。从模型预测的角度来看，ICL可以覆盖大部分微调的正确行为。

$$\frac{N_{(\mathrm{FT}>\mathrm{ZSL})\wedge(\mathrm{ICL}>\mathrm{ZSL})}}{N_{\mathrm{FT}>\mathrm{ZSL}}}$$

这些结果表明，从模型预测的角度来看，ICL 可以覆盖大部分微调的正确行为

基于Transformer注意力对偶形式的动量注意力

- 受注意力和梯度下降之间的对偶形式观点的启发，作者研究了是否可以利用动量来提高Transformer注意力。
- 动量法是梯度下降算法的一种改进，它引入了动量的概念以加速目标函数收敛过程并减小震荡。动量法的基本思想是在更新参数的过程中，不仅考虑当前的梯度方向，同时也考虑历史累积的梯度信息。



Model	SST5	IMDB	MR	CB	ARC-E	PIQA	Average
Transformer	25.3	64.0	61.2	43.9	48.2	68.7	51.9
$\text{Transformer}_{\text{MoAttn}}$	**27.4**	**70.3**	**64.8**	**46.8**	**50.0**	**69.0**	**54.7**

总结

- 当前仅考虑基于transformer的上下文学习，因为Transformer是当前NLP的主流架构。然而，至于上下文学习本身，弄清楚它如何在其他架构中工作也是一个有意义的问题。
- 虽然分类是上下文学习的一个典型应用，但本文没有考虑其他任务，可以在将来进行研究。