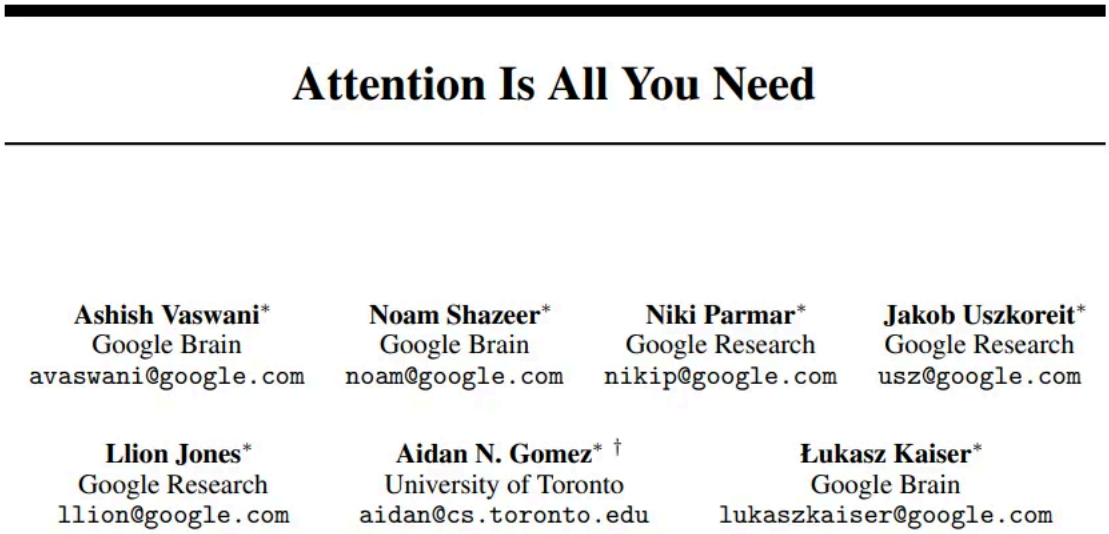
终于有人将Transformer可视化了！

Python开发者 2024年08月21日 10:13 浙江

都 2024 年，还有人不了解 Transformer 工作原理吗？快来试一试这个交互式工具吧。

2017 年，谷歌在论文《Attention is all you need》中提出了 Transformer，成为了深度学习领域的重大突破。该论文的引用数已经将近 13 万，后来的 GPT 家族所有模型也都是基于 Transformer 架构，可见其影响之广。

作为一种神经网络架构，Transformer 在从文本到视觉的多样任务中广受欢迎，尤其是在当前火热的 AI 聊天机器人领域。



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

不过，对于很多非专业人士来说，Transformer 的内部工作原理仍然不透明，阻碍了他们的理解和参与进来。因此，揭开这一架构的神秘面纱尤其必要。但很多博客、视频教程和 3D 可视化往往强调数学的复杂性和模型实现，可能会让初学者无所适从。同时为 AI 从业者设计的可视化工作侧重于神经元和层级可解释性，对于非专业人士来说具有挑战性。

因此，佐治亚理工学院和 IBM 研究院的几位研究者开发了**一款基于 web 的开源交互式可视化工具「Transformer Explainer」，帮助非专业人士了解 Transformer 的高级模型结构和低级数学运算**。如下图 1 所示。

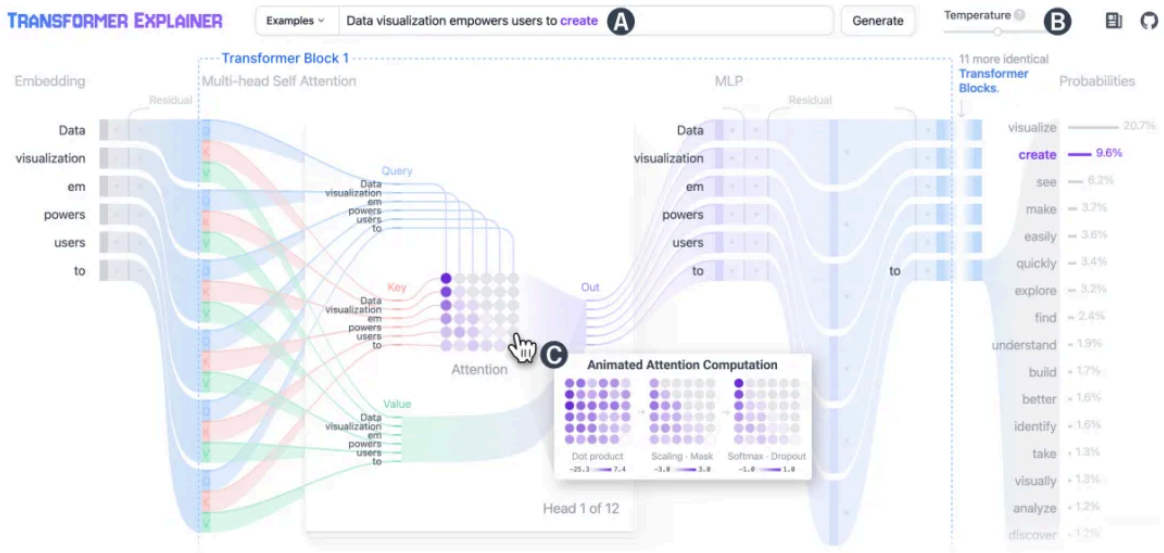


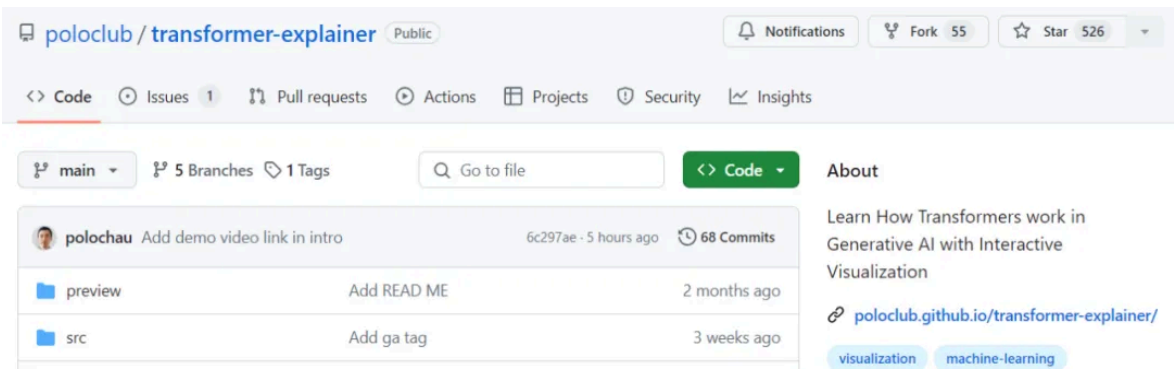
Figure 1: TRANSFORMER EXPLAINER helps users (A) visually examine how a text-generative Transformer model (GPT-2) transforms input text to predict next tokens, (B) interactively experiment in real time with key model parameters like *temperature* to understand prediction determinism, and (C) transition seamlessly between abstraction levels to visualize the interplay between low-level mathematical operations and high-level model structures.

Transformer Explainer 通过文本生成来解释 Transformer 内部工作原理，采用了**桑基图可视化设计**，灵感来自最近将 Transformer 视为动态系统的工作，强调了输入数据如何流经模型组件。从结果来看，桑基图有效地说明了信息如何在模型中传递，并展示了输入如何通过 Transformer 操作进行处理和变换。

在内容上，Transformer Explainer 紧密集成了对 Transformer 结构进行总结的模型概述，并允许用户在多个抽象层级之间平滑过渡，以可视化低级数学运算和高级模型结构之间的相互作用，帮助他们全面理解 Transformer 中的复杂概念。

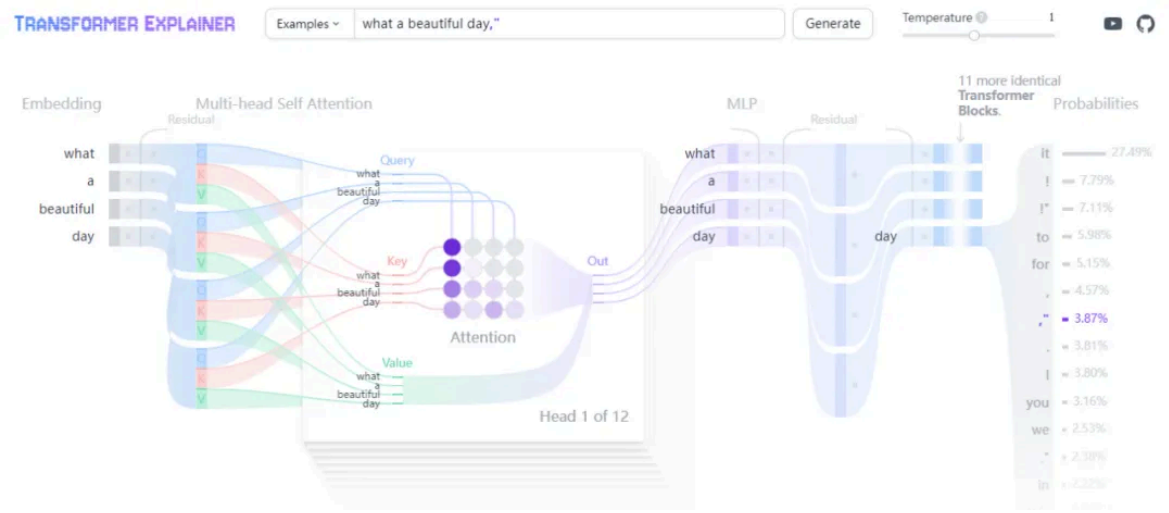
在功能上，Transformer Explainer 在提供基于 web 的实现之外，还具有实时推理的功能。与现有很多需要自定义软件安装或缺乏推理功能的工具不同，它集成了一个实时 GPT-2 模型，使用现代前端框架在浏览器本地运行。用户可以交互式地试验自己的输入文本，并实时观察 Transformer 内部组件和参数如何协同工作以预测下一个 token。

在意义上，Transformer Explainer 拓展了对现代生成式 AI 技术的访问，且不需要高级计算资源、安装或编程技能。而之所以选择 GPT-2，是因为该模型知名度高、推理速度快，并且与 GPT-3、GPT-4 等更高级的模型在架构上相似。

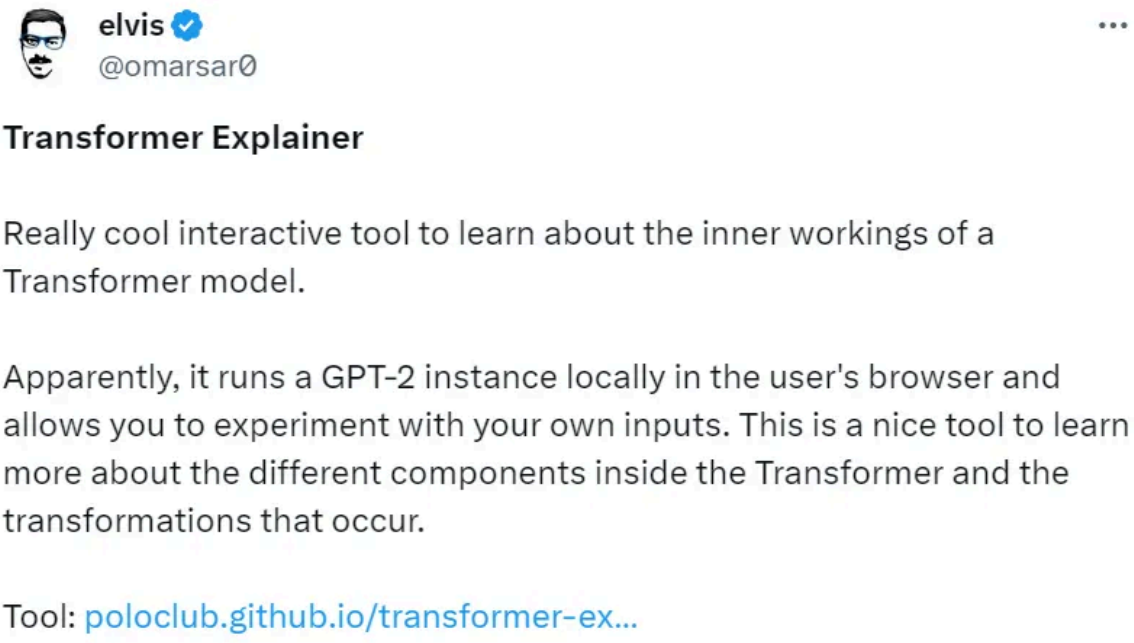


- 论文地址：https://arxiv.org/pdf/2408.04619
- GitHub 地址：http://poloclub.github.io/transformer-explainer/
- 类LLM可视化在线体验地址：https://t.co/jyBlJTMa7m

既然支持自己输入，试用了一下「what a beautiful day」，运行结果如下图所示。



对于 Transformer Explainer，一众网友给出了很高的评价。有人表示，这是非常酷的交互式工具。



有人称自己一直在等待一个直观的工具来解释自注意力和位置编码，就是 Transformer Explainer 了。它会是一个改变游戏规则的工具。

还有人展示了类LLM可视化中文项目。

展示地址：http://llm-viz-cn.iiiai.com/llm

这里不禁想到了另一位科普界的大牛 Karpathy，它之前写了很多关于复现 GPT-2 的教程，包括「纯 C 语言手搓 GPT-2，前 OpenAI、特斯拉高管新项目火了」、「Karpathy 最新四小时视频教程：从零复现 GPT-2，通宵运行即搞定」等。如今有了 Transformer 内部原理可视化工具，看起来两者搭配使用，学习效果会更佳。

Transformer Explainer 系统设计与实现

Transformer Explainer 可视化展示了基于 Transformer 的 GPT-2 模型经过训练是如何处理文本输入并预测下一个 token 的。前端使用了 Svelte 和 D3 实现交互式可视化，后端则利用 ONNX runtime 和 HuggingFace 的 Transformers 库在浏览器中运行 GPT-2 模型。

设计 Transformer Explainer 的过程中，一个主要的挑战是如何管理底层架构的复杂性，因为同时展示所有细节会让人抓不住重点。为了解决这个问题，研究者十分注意两个关键的设计原则。

首先，研究者通过多级抽象来降低复杂性。他们将工具进行结构化设计，以不同的抽象层次呈现信息。这让用户能够从高层概览开始，并根据需要逐步深入了解细节，从而避免信息过载。在最高层，工具展示了完整的处理流程：从接收用户提供的文本作为输入（图 1A），将其嵌入，经过多个 Transformer 块处理，再到使用处理后的数据来对最有可能的下一个 token 预测进行排序。

中间操作，如注意力矩阵的计算（图 1C），这在默认情况下被折叠起来，以便直观地显示计算结果的重要性，用户可以选择展开，通过动画序列查看其推导过程。研究者采用了一致的视觉语言，比如堆叠注意力头和折叠重复的 Transformer 块，以帮助用户识别架构中的重复模式，同时保持数据的端到端流程。

其次，研究者通过交互性增强理解和参与。温度参数在控制 Transformer 的输出概率分布中至关重要，它会影响下一个 token 预测的确定性（低温时）或随机性（高温时）。但是现有关于

Transformers 的教育资源往往忽视了这一方面。用户现在能够使用这个新工具实时调整温度参数（图 1B），并可视化其在控制预测确定性中的关键作用（图 2）。

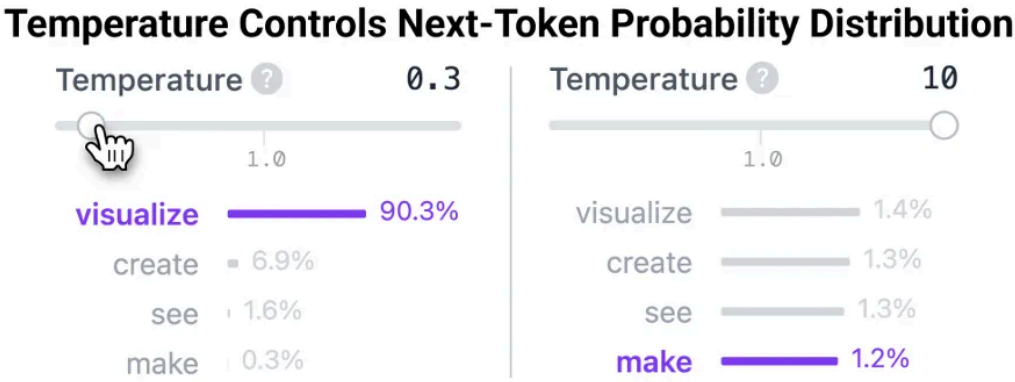


Figure 2: The temperature slider lets users interactively experiment with the temperature parameter's impact on the next token's probability distribution. **Left**: lower temperatures sharpen the distribution, making outputs more predictable. **Right**: higher temperatures smooth the distribution, resulting in less predictable outputs.

此外，用户可以从提供的示例中选择或输入自己的文本（图 1A）。支持自定义输入文本可以让用户更深入参与，通过分析模型在不同条件下的行为，并根据不同的文本输入对自己的假设进行交互式测试，增强了用户的参与感。

那在实际中有哪些应用场景呢？

Rousseau 教授正在对自然语言处理课程的课程内容进行现代化改造，以突出生成式 AI 的最新进展。她注意到，一些学生将基于 Transformer 的模型视为捉摸不透的「魔法」，而另一些学生则希望了解这些模型的工作原理，但不确定从何入手。

为了解决这一问题，她引导学生使用 Transformer Explainer，该工具提供了 Transformer 的互动概览（图 1），鼓励学生积极进行实验和学习。她的班级有 300 多名学生，而 Transformer Explainer 能够完全在学生的浏览器中运行，无需安装软件或特殊硬件，这是一个显著的优势，消除了学生对管理软件或硬件设置的担忧。

该工具通过动画和互动的可逆抽象（图 1C），向学生介绍了复杂的数学运算，如注意力计算。这种方法帮助学生既获得了对操作的高层次理解，又能深入了解产生这些结果的底层细节。

Rousseau 教授还意识到，Transformer 的技术能力和局限性有时会被拟人化（例如，将温度参数视为「创造力」控制）。通过鼓励学生实验温度滑块（图 1B），她向学生展示了温度实际上是如何修改下一个词元的概率分布（图 2），从而控制预测的随机性，在确定性和更具创造性的输出之间取得平衡。

此外，当系统可视化 token 处理流程时，学生们可以看到这里并没有任何所谓的「魔法」—— 无论输入文本是什么（图 1A），模型都遵循一个定义明确的操作顺序，使用 Transformer 架构，一次只采样一个 token，然后重复这一过程。

未来工作

研究者们正在增强工具的交互式解释来改善学习体验。同时，他们还在通过 WebGPU 提升推理速度，并通过压缩技术来减小模型的大小。他们还计划进行用户研究，来评估 Transformer Explainer 的效能和可用性，观察 AI 新手、学生、教育者和从业者如何使用该工具，并收集他们希望支持的额外功能的反馈意见。

还在等什么，你也上手体验一下，打破对 Transformer 的「魔法」幻想，真正了解这背后的原理吧。

- 论文地址：https://arxiv.org/pdf/2408.04619
- GitHub 地址：http://poloclub.github.io/transformer-explainer/
- 类LLM可视化在线体验地址：https://t.co/jyBlJTMa7m
