arXiv:2305.03210v2 [cs.HC] 9 Aug 2023

AttentionViz：Transformer注意力的全局视图

凯瑟琳 ne Yeh、Yida Chen、Aoyu Wu、Cynthia Chen、Fernanda Viégas 和 Martin Wattenber g

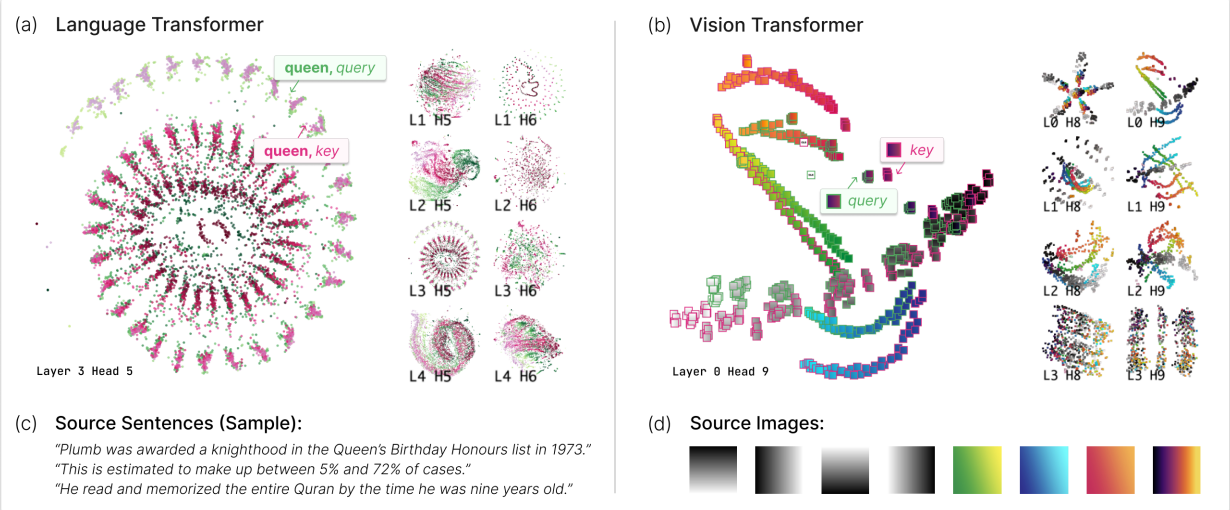


图 1：AttentionViz 是我们的交互式可视化工具，它通过为 查询 和 键 创建联合嵌入空间，使用户能够大规模探索 Transformer 自注意力。 **(a)** 在语言 Transformer 中，这些可视化揭示了引人注目的视觉痕迹，可与注意力模式联系起来。散点图中的每个点代表一个词的查询或键版本，具体由点的颜色表示。用户可以探索单个注意力头（左），也可以缩小视图，查看注意力的"全局"概况（右）。 **(b)** 我们的可视化还揭示了视觉 Transformer 中有趣的现象，例如按色调和亮度对图像块进行分组的注意力头。边框颜色表示图像块的查询嵌入（绿色）或键嵌入（粉色）。 **(c)**提供示例输入句子（来自 [23]）和 **(d)** 图像（合成数据集）供参考。

摘要—Transformer 模型正在推动机器学习领域的变革，但其内部运作机制仍然神秘莫测。在这项工作中，我们提出了一种新的可视化技术，旨在帮助研究人员理解*Transformer* 中的自注意力机制，该机制使这些模型能够学习序列中各元素之间丰富的上下文关系。我们方法的核心思想是对查询与键向量进行联合嵌入可视化；Transformer 模型利用这些向量来计算注意力。与以往的注意力可视化技术不同，我们的方法能够分析多个输入序列中的全局模式。我们基于这些联合查询-键嵌入创建了交互式可视化工具 AttentionViz（演示地址：http://attentionviz.com），并利用它研究语言和视觉 Transformer 中的注意力机制。通过多个应用场景和专家反馈，我们展示了该方法在增进模型理解以及揭示查询-键交互新见解方面的价值。

索引词—Transformer、注意力、自然语言处理、计算机视觉、可视分析

1 引言

Transformer 神经网络架构 [52]正在对从自然语言处理（NLP）[13,42]到计算机视觉 [14]等领域产生重大影响。事实上，Transformer 已应用于大型现实世界系统，数亿人都在使用这些系统（例如，Stable Diffusion、ChatGPT、Microsoft Copilot）。然而，这一成功背后的机制仍在很大程度上未被揭示，尤其是随着模型复杂度和规模不断增长，新能力也持续涌现 [11,60]。深入理解 Transformer 模型，有助于我们构建更可靠的系统、排查问题并探索改进方向。

在这项工作中，我们介绍了一种新的可视化技术，旨在更好地理解 Transformer 的运作方式。（我们在第 2 节简要介绍 Transformer。）我们的分析对象是 Transformer 典型的自注意力 机制，它使这些模型能够学习并利用输入元素之间丰富多样的关系。尽管注意力模式已得到深入研究，但以往的技术通常只可视化与单个输入序列（例如一个句子或一幅图像）相关的信息。典型方法会为给定的输入序列创建注意力权重的二部图 [51,53] 或热力图 [20,30]表示。

我们的方法提供了一个更高层次的视角，使我们能够同时查看许多输入序列的自注意力模式。这一方法的灵感之一来自 Activation Atlas [5], 等工具的成功。Activation Atlas 允许研究者"缩小视野"以概览神经网络，然后再深入查看细节。在我们的案例中，我们希望构建一种"注意力图谱"，为研究者提供丰富而详细的视图，以了解 Transformer 的各个注意力头如何运作。其主要的新技术是将 Transformer 使用的 查询向量和 键向量的联合嵌入进行可视化，从而为单个注意力头创建一种视觉特征。

为展示我们的方法，我们实现了 AttentionViz，这是一个交互式

• 作者就职于哈佛大学。*Viégas* 和 *Wattenberg* 也就职于 *Google*，但这项工作是在哈佛完成的。电子邮件：*{catherineyeh, yidachen,aoyuwu, fernanda, wattenberg}@g.harvard.edu, cynthiachen@college.harvard.edu*

稿件于 *201x* 年 *xx* 月 *xx* 日收到；于 *201x* 年 *xx* 月 *xx* 日录用。发布日期：*201x* 年 *xx* 月 *xx* 日；当前版本日期：*201x* 年 *xx* 月 *xx* 日。有关获取本文重印本的信息，请发送电子邮件至：*reprints@ieee.org*。数字对象唯一标识符：*xx.xxxx/TVCG.201x.xxxxxxx*

可视化工具，允许用户探索语言和视觉 Transformer 中的注意力。AttentionViz 支持通过多个细节层级进行探索（图 1），既提供全局视图，让用户一览所有注意力头，也能放大查看单个注意力头或输入序列的细节。

我们通过若干使用 AttentionViz 的应用场景以及领域专家访谈，展示了该技术的实用价值。为具体说明，我们聚焦于可视化能够揭示的几种广泛使用的 Transformer：BERT [13], GPT-2 [41], 和 ViT [14]。我们发现了 BERT 中与注意力模式相关的若干可识别的“视觉痕迹”，检测到 ViT 视觉注意力机制中新颖的色调/频率行为，并揭示了 GPT-2 中可能存在的异常行为。用户反馈也支持了我们的方法在大规模可视化其他嵌入方面具有更广泛的适用性。

总而言之，本研究的贡献包括：

- 一种基于联合查询-键嵌入、用于探索 Transformer 模型中注意力趋势的可视化技术。
- AttentionViz 是一款交互式工具，它运用我们的技术，以多个尺度研究视觉和语言 Transformer 中的自注意力。
- 应用场景和专家反馈展示了 AttentionViz 如何揭示有关 Transformer 注意力模式的见解。

2 Transformer 模型背景

Transformer（*Transformer*）于 [52], 提出，是一种用于处理序列输入的神经网络架构。本文不打算全面介绍 Transformer，但要理解我们的工作，有几个关键概念需要了解。首先，Transformer 接收一组向量作为输入（通常称为**嵌入**）。嵌入可以表示多种输入类型。在基于文本的 Transformer 中，它们对应单词或单词片段；在视觉 Transformer 中，它们对像素图像块进行编码。

网络通过一系列**注意力层**迭代地变换这些向量，每一层都会在成对的嵌入之间传递信息。“注意力”这个名称表明，并非所有嵌入之间的关联都同样紧密；某些嵌入对之间的交互会更强，也就是说，它们会彼此“关注”得更多。注意力层决定哪些嵌入对应该进行交互，以及它们之间应该传递哪些信息。

例如，在一个处理句子“The brown capybara is sleeping now”中各单词的 Transformer 里，人们可能会预期，“capybara”和“is”对应的嵌入之间有较高的注意力（以及信息流动），而“brown”和“now”之间则不会。本文重点关注的**自注意力** 机制，使 Transformer 能够学习并利用序列元素之间丰富的关系，从而在各种自然语言处理和计算机视觉任务中显著提升性能 [13, 14, 41]。

嵌入向量对彼此产生注意力可能有不同的原因。例如，在我们的示例句子中，“brown”和“capybara”之间存在形容词-名词关系，而“capybara”和“is”之间则构成主谓关系。为了支持多种关系类型，Transformer注意力层由多个**注意力头**组成，每个注意力头都能表示不同的注意力模式和信息流。

每个注意力头都使用由查询权重矩阵W_Q 和键权重矩阵W_K计算出的双线性形式，来计算自身的注意力模式。具体而言，对于两个嵌入向量x 和y，注意力与**查询向量**$W_Q x$和**键向量**$W_K y$的缩放内积有关。设d 为$W_K y$的维度，则有：

$$f(x,y) = \frac{1}{\sqrt{d}} \langle W_Q x, W_K y \rangle$$

给定嵌入向量 $\{x_1, x2, ..., x_n\}$，我们使用 softmax 函数计算x_i 与其他向量之间的自注意力：

$$attn(x_i, x_j) = \text{softmax}_j(f(x_i,x), \ldots, f(x_i,x_n)) = e^{f(x_i,x_j)} / \sum_k e^{f(x_i,x_k)}$$

值得注意的是，这个公式表明，查询向量和键向量之间的点积越大，最终的注意力值就越高；这一事实正是我们在联合嵌入可视化中所依据的。

Transformer 架构远不止我们在此介绍的内容。具体而言，我们只描述了嵌入两两之间的注意力权重，而没有说明它们之间流动的具体信息。（正如后文所述，这一领域仍有很大的进一步研究空间。）不过，还有最后一个技术要点值得一提，因为它有助于解读本文后面的图像。Transformer 接收的初始嵌入通常包含一个表示其顺序（对于一维序列）或空间布局（对于网格，例如视觉 Transformer）的向量。对于序列，这些位置向量由三角函数定义，并位于高维空间中类似螺旋线的曲线上（见 [52]）。

2.1 本文研究的模型

我们研究了三种 Transformer 模型：BERT（语言）、GPT-2（语言）和 ViT（视觉）。它们都是机器学习领域的重要研究对象，且涵盖了多种 Transformer 架构和应用。**BERT**，即来自 Transformer 的双向编码器表示 [13], ，是一种多层 Transformer 编码器。作为双向模型，BERT 可以双向关注词元（即输入元素）。**GPT-2**，即生成式预训练 Transformer 2 [42], ，是一种多层 Transformer 解码器。GPT-2 是单向模型，也就是说，它只关注前面的词元。**ViT**，即视觉 Transformer [14], ，采用基于自注意力的 Transformer 架构，将图像拆分为“图像块”，并将它们视为句子中的词元。与 BERT 类似，ViT 是一种多层双向 Transformer 编码器。在这项工作中，我们考察了 ViT 在 16x16（**ViT-16**）和 32x32（**ViT-32**）图像块尺寸下的表现。

3 相关工作

许多研究人员尝试探究 Transformer 的内部运作机制。 [8, 34] 通过探索模型学到的语言表征，试图理解基于 Transformer 的语言模型性能提升的原因； [49] 则发现，BERT 重现了自然语言分析中的经典步骤，从词性标注到关系分类。另一种广受欢迎的方法是机制可解释性，即对 Transformer 模型进行逆向工程（例如， [15, 16, 37]）。作为 Transformer 基石的注意力机制也得到了深入研究。例如，在自然语言处理系统中，注意力似乎与句法结构相关 [9, 57] ；而在视觉变换器中，它似乎与格式塔式分组相关 [33]。研究人员还将 ViT 的视觉注意力机制与卷积滤波器进行了比较，发现注意力对于图像遮挡、损坏和高频噪声具有更强的鲁棒性 [35, 40]。在讨论相关工作时，我们着重介绍用于研究 Transformer 注意力的视觉方法。

3.1 可视化单个输入序列中的注意力

无论是在语言还是视觉 Transformer [4, 12, 21, 31, 39] 中，注意力模式都天然适合进行可视化。这些可视化主要侧重于展示 单个输入序列中 查询词元与键词元之间的注意力权重，通常采用二部图（例如，[48, 51, 53]）或热力图（例如，[1, 10, 20, 22, 25, 30, 43]）的形式。[50,

已有一些可视化方法能够支持跨多个模型或层进行比较。例如，Attention Flows [12]支持用户比较 BERT 的层内和层间注意力，以及给定单个句子时不同注意力头之间的注意力。Dodrio [59]采用应用于单个输入的网格视图，实现了注意力头之间的直接比较。另一个系统 VisQA [21], 通过展示语言自注意力、视觉自注意力和语言-视觉交叉注意力的热图，可视化视觉问答任务中不同注意力头的注意力。然而，即使在这些模型比较系统中，分析人员也必须逐个查看不同的输入，一次一个，以识别并验证给定注意力头的模式。

3.2 超越单一输入：嵌入可视化与激活最大化

自然地，我们会寻求适用于多个输入的模式。为了实现这一目标，一种行之有效的技术是将多个输入序列中的嵌入向量集合可视化。

[3, 18, 19, 46, 47, 58]。例如， [43] 将同一个词在许多不同语境中的 BERT 嵌入可视化，并发现了与词义对应的簇。在对句法处理的探索中， [8] 将多语言 BERT 模型中的嵌入可视化，再次发现了有助于解读的有意义簇。LMFingerprints [45] 使用基于树的径向布局，比较不同语言模型中的嵌入向量。

第二种用于视觉 Transformer 的技术 [17, 61], 旨在找出能够最大化特定单元激活值的图像。将该技术应用于嵌入向量时，可以得到清晰且易于解释的结果。不过，作者指出，将其应用于查询向量和键向量时，似乎无法产生有用的结果。

3.3 文献中的空白

我们注意到三个 gap空白，这些空白g 促使我们开展这项工作。

首先，已有研究表明，可视化嵌入向量是分析多个输入之间模式的一种有效技术，但据我们所知，尚无研究系统地尝试可视化 Transformer 模型中的查询和键嵌入。 [6] 还指出，自注意力的中间产物（例如查询和键）尚未得到充分探索。这些观察促使我们提出联合查询-键嵌入技术。

其次，尽管已有可视化技术被提出用于比较多个嵌入（例如，[2, 3, 26]），但这些方法通常只能处理少量嵌入，无法满足我们比较不同 Transformer 注意力头和层中的嵌入的需求。因此，我们设计了一种全局矩阵视图，以大规模可视化查询-键嵌入。

最后，二部图表示已被证明有助于分析基于自然语言处理的 Transformer，但我们尚未见过它们应用于视觉任务。我们通过创建二部图风格可视化，研究 ViT 中的图像注意力模式，探索这一方向。

4 目标与任务

本研究的总体目标是设计一种新颖的可视化技术，以便探索 Transformer 模型中的全局注意力趋势。为了收集有关这一构想的初步反馈，并进一步了解用户需求，我们与 5 位对模型可解释性感兴趣的机器学习（ML）研究人员进行了交流（其中包括 4 名博士生和 1 名教授）。在这些个人访谈中，我们请专家介绍他们在使用 Transformer 时的当前实践和面临的挑战，以及注意力可视化如何帮助他们实现研究目标。我们将这些专家称为**E1-5**。

4.1 目标

最终，我们与专家的交流归纳出了三个主要目标：

G1 - 了解自注意力如何影响模型行为。总体而言，5 位专家都希望更好地理解不同注意力头的行为，以及 Transformer 模型通过其特有的自注意力机制学到了什么。因此，他们表示希望能够快速、轻松地探索注意力模式。**E2**解释说：“注意力仍然像个黑箱，存在许多未解之谜”，因此，深入了解 Transformer 注意力模式或许能让我们明白“为什么大型语言模型在推理任务和数学方面会失败”，等等。

G2 - 比较和对照注意力头。**E5** 提到，将注意力头之间的差异可视化有助于假设生成，这是其研究流程的第一步：“可视化有助于提出待检验的假设，并直观地了解 Transformer 在做什么。”此外，3 位专家（**E1, E2, E5**）指出，比较注意力头有助于模型剪枝和编辑。也就是说，如果两个注意力头的行为看起来相似，也许可以移除其中一个，而不会显著影响模型性能。用**E1**的话来说，比较注意力头或许能帮助我们“找到模型中真正有用的部分”。

G3 - 识别注意力异常。四位研究人员（**E2-5**）希望通过探索注意力模式，识别变换器中的不规则现象和潜在行为问题。这些信息随后可用于模型调试。例如，**E4** 表示：“将注意力可视化，有助于你注意到模型何时关注了错误的对象，即使结果是正确的。”**E3** 对此表示赞同，并重申了调试的重要性，尤其是在模型训练的背景下：“训练经常失败并终止，但很难理解它为什么失败，或为什么会产生意外行为。”

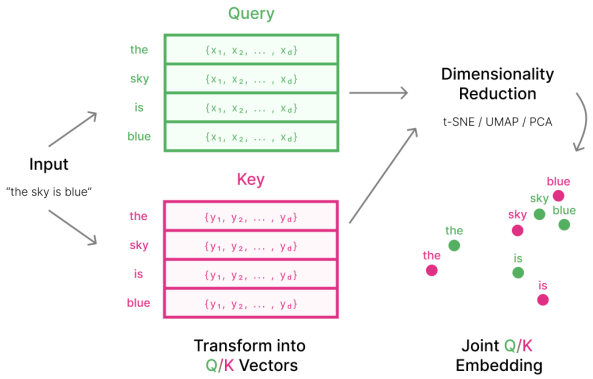


图 2：为单个注意力头创建联合查询-键嵌入空间。在自然语言处理场景中，给定一个输入句子，我们首先将每个词元转换为相应的查询向量和键向量。然后，使用 t-SNE/UMAP/PCA 将这些 $1 \times d$向量投影为二维/三维散点图坐标。对于所使用的 BERT、GPT-2 和 ViT 模型，$d = 64$。

调试的重要性，尤其是在模型训练的背景下：“训练经常失败并终止，但很难理解它为什么失败，或为什么会产生意外行为。”

4.2 任务

基于这些目标，我们制定了以下一组设计任务：

T1 - 大规模可视化注意力头。为了帮助用户快速探索模型行为 **[G1]** 并轻松比较和对照注意力模式 **[G2]**, , 我们的工具会同时可视化各 Transformer 层中的自注意力头。

T2 - 探索查询-键交互。**E1** 和 **E4** 表示希望更好地理解查询-键配对信息，以加深对 Transformer 自注意力的理解。因此，我们的工具还通过可视化查询-键交互来支持注意力模式比较 **[G2]** 和异常检测 **[G3]** 。

T3 - 在多个层级探查注意力。我们的工具通过在句子/图像、注意力头和模型层级提供可视化，实现注意力的局部和全局比较 **[G2]**。在单一界面中灵活切换多个视图，也有助于知识发现 **[G1]**，并帮助用户识别模型异常 **[G3]**。

T4 - 自定义模型和数据输入。 AttentionViz 可轻松扩展以支持新的 Transformer 和数据集，从而快速进行视觉比较 **[G2]** 并综合分析注意力模式 **[G1]**，涵盖不同模型和模态（语言与视觉）。

5 查询/键嵌入与 AttentionViz 的设计

为实现这些目标和任务，我们构建了一个名为 AttentionViz 的工具。该工具采用的主要技术是可视化每个注意力头的查询向量和键向量的联合嵌入。本节首先介绍这一技术背后的动机和数学原理，然后讨论完整应用的设计。

5.1 查询/键嵌入可视化

AttentionViz 背后的技术相对简单，不过正如我们下面将要说明的，要使其有效，需要运用两种数学技巧。回想一下，每个 Transformer 注意力头都会分别通过矩阵W_Q 和 W_K将输入嵌入转换为查询向量和键向量（第 2 节）。这些矩阵会将原始向量嵌入投影到一个较低维度的空间，本质上是从高维向量嵌入中选取特定类型的信息。因此，通过检查查询向量和键向量，我们或许可以了解W_Q和 W_K所选取的信息。

一个核心观察是，查询向量和键向量的相对位置可以提供线索，帮助判断注意力将如何分布，因为注意力系数取决于查询和键之间的点积。要理解其中的原因，不妨考虑这样一种假设情形：查询向量和键向量的范数始终相同。此时，距离越近，便会.....

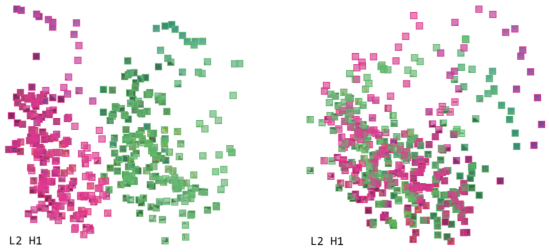


图 3：**左：** 联合嵌入空间中的原始查询和键。**右：** 将键平移以对齐查询和键的质心后，重叠程度增加。

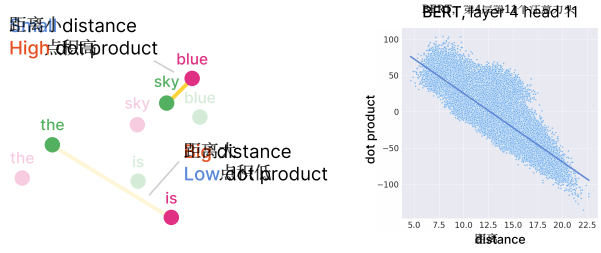


图 4：**(a)** 理想的距离与注意力关系，其中点积较高的查询-键对在联合嵌入空间中距离更近。**(b)**BERT 中的一个注意力头示例，查询-键距离与点积之间呈现较强的负相关（**-0.983**）。

与较高的注意力系数直接相关。在实践中，查询向量和键向量的范数会发生变化，因此点积与距离之间的关系并不精确。不过，正如以下章节所述，我们可以使这种关系出乎意料地接近。

图 2 通过语言Transformer中单个注意力头的合成示例展示了我们的方法。要创建联合嵌入，我们首先获取给定句子中每个词元的查询和键向量表示（第 2 节）。然后，我们使用三种降维方法之一，将这些高维向量投影到共享的低维子空间： **t-SNE** [50], **UMAP** [32], 或 **主成分分析** [24]。这些降维算法的输出是二维/三维散点图，其中每个点代表一个查询或键词元。同样的流程也可用于为 ViT 注意力头创建联合嵌入，此时每个词元都是一个图像块。默认情况下，我们用 绿色 表示查询，用 粉色表示键。不过，用户可以选择多种颜色编码（参见第 5.2 节），也可以轻松地将其他调色板替换到系统中。

我们的联合嵌入允许用户探索查询与键之间的细粒度交互，而这些图的形状通常可以作为底层自注意力模式的视觉指标（见第7节）。对于给定的数据集，每种降维技术都会产生不同的模式，从而提供独特的洞见和使用场景。

5.1.1 向量归一化

在设计 AttentionViz 时，我们注意到两个“自由参数”，它们可以在不丢失任何信息的情况下进行调整。调整这两个参数能够使嵌入距离与注意力权重之间建立更紧密的联系，并显著提高可视化结果的可读性。这些归一化操作在降维之前进行（图 2）。

键平移： 在我们的可视化中，查询向量和键向量有时彼此分离得很远（图 3，左）。这种分离使得直接比较查询和键嵌入变得困难。不过，一个简单的数学技巧可以让我们将这些嵌入移得更近，同时不影响任何给定输入序列的自注意力计算。具体来说，请注意 softmax

函数具有平移不变性：即，对于任意常数 a，我们有 $\text{softmax}_j(x_1+a, x_2+a, \ldots) = \text{softmax}_j(x_1, x_2, \ldots)$。现在，考虑一个查询向量 x 和键向量 $y_1, \ldots, y_n$。对于任意向量 v，我们有：

$$\begin{aligned}\text{attention}_j(x) &= \text{softmax}_j(\langle x, y_1\rangle, \langle x, y_2\rangle, \ldots) \\ &= \text{softmax}_j(\langle x, y_1\rangle + \langle x, v\rangle, \langle x, y_2\rangle + \langle x, v\rangle, \ldots) \\ &= \text{softmax}_j(\langle x, y_1 + v\rangle, \langle x, y_2 + v\rangle, \ldots)\end{aligned}$$

其中，第二个等式由平移不变性得出。这意味着，在不改变任何给定输入的注意力模式的情况下，我们可以平移所有键向量，使得每个注意力头的查询分布和键分布具有相同的质心。这使得比较查询和键变得容易得多（图 3，右侧）。

缩放查询和键： 在一些 Transformer（例如 GPT-2）中，我们观察到平均查询范数与平均键范数相差很大的情况。这种差异使得难以在联合嵌入中解读键与查询之间的关系。从数学上看，这表明点积与距离之间的关系较差；从视觉上看，这意味着查询可能聚集成一个很小的簇，周围则散布着一片松散的键云。

幸运的是，缩放也是系统的另一个“自由参数”。自注意力水平仅取决于查询向量和键向量的点积，因此，如果我们将所有查询向量缩放 $c \neq 0$ 倍，并将所有键向量缩放 c^{-1} 倍，注意力值就不会改变。这使得高注意力的查询-键对能够在我们的联合可视化中彼此靠得更近，如图 4a 所示。（一个微妙之处：单独缩放不会改变余弦距离；但与平移归一化结合时，它会产生显著影响。）

为了确定 c 的最优值，我们可以定义一个 加权相关性 指标，对距离较小的查询-键对赋予更高的权重，因为在联合可视化中，我们最关注彼此邻近的查询和键。因此，我们可以选择一个缩放因子 c，使查询-键点积与距离之间的加权相关性达到最大。通过这种缩放方法，联合嵌入空间中的距离能够最准确地表示查询和键之间的实际注意力值。

5.1.2 距离作为注意力的代理指标

如上所述，理想情况下，如果一个查询-键对的点积较大且为正（对应较高的最终自注意力值），它们在嵌入空间中应彼此更接近，反之亦然（图4a）。因此，我们预期在联合查询-键嵌入中，距离与注意力呈负相关。我们通过计算 BERT、GPT-2 和 ViT 中每个注意力头的余弦距离与点积之间的斯皮尔曼秩相关系数，研究这种潜在关联。我们还尝试在创建查询和键的 t-SNE 与 UMAP 投影时使用欧几里得距离作为距离度量，但这通常会导致距离与点积之间的相关性较弱。

在多个数据集和模型中，距离与注意力之间的关系基本成立。例如，对于 Wiki-Auto 数据 [23], , 查询-键距离与点积之间的平均相关系数为**-0.938**（BERT）和**-0.792**（GPT）。图4b 展示了 BERT 的一个示例结果。在所使用的 COCO 图像集上 [28], , 平均相关系数为**-0.873**（ViT-32）和**-0.884**（ViT-16）。

5.2 颜色编码

为了可视化查询和键的不同属性，AttentionViz 提供了多种颜色编码。默认选项按照令牌类型为点着色，即查询或键。对于视觉 Transformer，用户可以按照图像块的行或列着色，以可视化位置模式（图10）。由于图像本身包含颜色信息，我们还允许用户查看原始图像块，而不添加其他样式元素（图8）。

对于语言变换器，我们支持两种位置颜色方案： 归一化 和 离散。要计算归一化位置，我们将每个词元在句子中的位置除以句子长度，从而生成连续颜色刻度。较浅的色调表示更接近句子开头的词元（图 5b）。我们的离散位置编码取每个词元的位置，并应用取模运算符，得到其除以 5 后的余数。因此， 1^{st} 和 6^{th} 词元会获得相同的颜色， 2^{nd} 和 7^{th} 词元会获得相同的颜色，依此类推。我们使用相同的五种颜色对不同位置的查询和键进行编码，其中前者使用较深的色调。尽管这种方案会引入

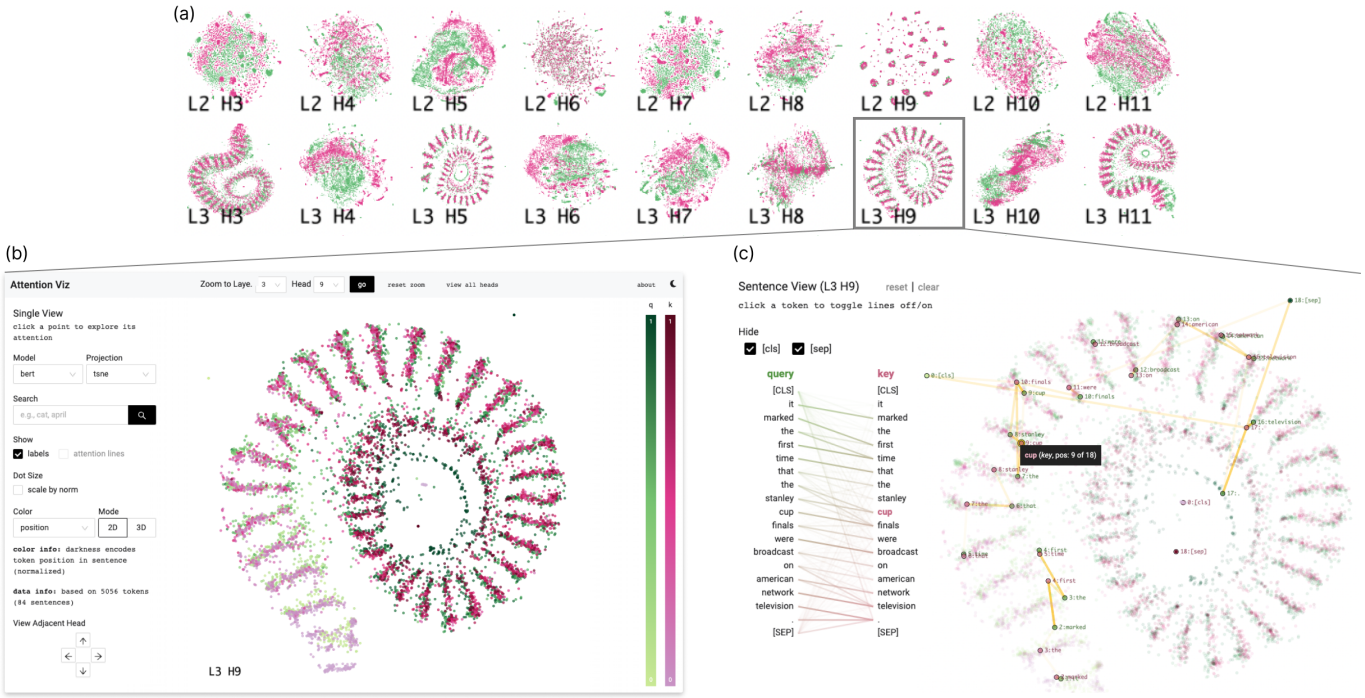


图5：连接 BERT 中的形式与功能。**(a)** 在矩阵视图中，第3层有多个螺旋形图。**(b)** 使用单视图放大其中一个注意力头（*L3 H9*）后，我们可以通过使用从浅到深的配色方案来查看位置注意力模式，该方案对输入序列中的位置进行编码。**(c)** 通过探索句子级可视化，可以验证这些模式。

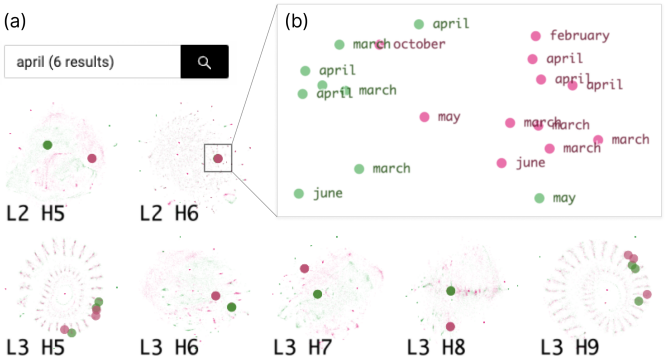


图6：使用全局搜索探索注意力模式。**(a)**搜索结果簇较少的注意力头通常表现出更多语义行为，而结果分散的注意力头则更关注词元位置。**(b)** 放大 L2 H6（一个具有单个主要结果簇的注意力头）后，我们确实看到了大量语义相关的查询和键令牌。

对于长度为 >5 的句子会产生歧义，但我们发现，离散着色有助于观察基于较小位置偏移的关系（例如，查询关注相距一步的键，如图 11左图所示）。通过将鼠标悬停在词元上也可以轻松消除歧义；在我们的探索中，没有发现偏移量大于 2 或 3 的模式。用户还可以按 查询/键范数 进行着色（图 12a）。

5.3 视图

AttentionViz 提供了三种主要的交互式注意力探索视图：矩阵视图、单视图和句子/图像视图。

5.3.1 矩阵视图

AttentionViz 中的初始视图是矩阵视图，它使用小多图一次性可视化 Transformer 中的所有注意力头（图5a），直接回应了 **[T1]** 和 **[T3]**。每一行对应一个模型层，界面顶部显示较早的层底部则显示较晚的层。借助这种“全局”视角，与单图（例如 [53]）或实例级可视化（例如 [3,47]）相比，用户可以更轻松地扫描不同 Transformer 层和注意力头中的模式。本研究使用的所有模型架构都相同：12 层，每层 12 个注意力头 $=144$，总计 144 个注意力头，但我们的系统也能扩展到其他维度。

在矩阵视图中，用户可以查看使用 t-SNE、UMAP 或主成分分析创建的联合查询-键嵌入。用户还可以在不同的模型类型（即 BERT、GPT-2、ViT-16/32）或数据集之间切换 **[T4]**,探索不同的配色方案，并以二维或三维形式查看生成的图。矩阵视图支持全局搜索功能（图6a），该功能有助于突出显示不同注意力头中的词元位置模式，并提供另一种大规模分析注意力的方法（见第7节）。

5.3.2 单视图

用户可以点击矩阵视图中的任意图表，放大查看单视图(图5b)，从而更细致地探索单个注意力头 **[T3]**。与矩阵视图一样，用户可以在单视图中切换颜色方案、维度、投影模式、数据集和模型**[T4]**；所有图形变化都会在各视图间同步，以便进行比较。用户可以点击某个点，高亮显示对应输入序列中的所有词元，从而突出显示联合嵌入空间中相关的查询和键。用户还可以选择将注意力连线投影到散点图上，以连接查询和键词元（图5c）。为提高可读性，我们仅显示每个词元对应的前 2 个注意力权重。我们的注意力连线功能支持 **[T2]**，并为在头级别可视化 Transformer 中的注意力模式提供了一种新方式。

在单视图中，用户还可以搜索词元并使用我们的标注功能，发现数据中的语义模式，类似于 [47]。例如，在图6b 中，搜索结果显示，对于这个 BERT 注意力头，含义相似的查询/键词元在联合嵌入中彼此相邻，表明它们之间存在较强的注意力（第5.1.2节）。

5.3.3 句子/图像视图

句子/图像视图 允许探索单个句子或图像中的细粒度注意力模式 **[T2, T3]**。这两种视图都可以】【。

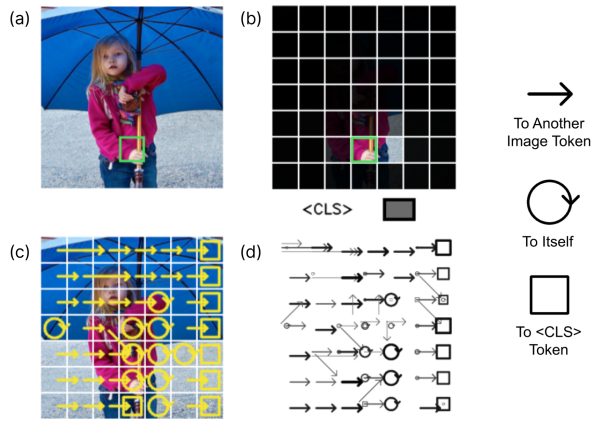


Fig. 7：ViT 图像视图。（a）原图像。（b）我们将鼠标悬停在一个图像块上，并用绿色边框突出显示所选词元（查询）。（c）叠加的注意力箭头。（d）全局注意力流。方形图标表示某个图像块与 <CLS> 词元之间具有最强的注意力连接，该词元不在原始图像中。

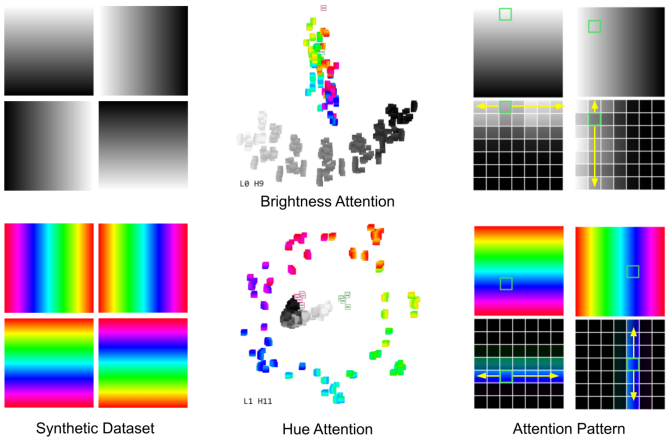


Fig. 8：在 ViT-32 矩阵视图中，我们发现了两个有趣的视觉注意力头：其中一个注意力头根据亮度对黑白图像令牌进行排序，另一个则根据色调对彩色图像块进行排列。图像视图中显示的注意力模式证实了具有相同亮度的图像块之间存在注意力流。

与单视图同步，在每个查询/键散点图上叠加的注意力连线保持一致，带来流畅的用户体验。

句子视图。使用 BERT 或 GPT-2 时，用户可以点击单视图中的某个点，在左侧边栏中打开句子视图。该视图显示受 BertViz 启发的句子级注意力可视化，并突出显示所点击的词元 [53]（图 5c）。我们也考虑过使用热力图可视化（例如 [39]），但对于较长的句子而言，二部图方法似乎具有更好的可读性，也更便于探索模式。连接左列查询词元与右列键词元的连线，其透明度表示相应的注意力强度。将鼠标悬停在某个词元上时，会突出显示该词元特有的注意力连线。为减少 BERT 中分类词元和分隔符，或 GPT-2 中首个词元（第 7节）带来的噪声，用户可以隐藏来自这些特殊词元的注意力连线。其他查询词元和键词元也可以切换显示或隐藏，所有注意力连线将相应地重新归一化。用户还可以选择查看每个注意力头的聚合注意力模式，从而提供另一层比较维度（图 11a）。

图像视图。对于 ViT 中基于图像的输入，当用户点击一个图像块时，侧边面板会显示对应的原始图像，并用彩色边框突出显示被点击的词元（图。7a）。用户还可以看到叠加了注意力热图的图像，其中透明度表示被点击的图像块与图像其他区域之间的注意力权重（图 7b）。

除了可视化单个词元的注意力外，图像视图还允许用户通过展示不同图像块之间带箭头的注意力连线，探索图像中的整体注意力模式。可视化注意力箭头时，我们为用户提供了两种选择。第一种是在原始图像块上叠加箭头，每个箭头表示起始图像块与目标图像块之间最强的注意力连接（图 7c）。这会生成一个简化的二分注意力图，帮助用户描述特定注意力头中最重要的模式。第二种是在原始图像旁展示所有较强的注意力连接（即 $attn(x_i,xj) > 0.1$），以便更全面地查看注意力（图 7d）。在此可视化中，我们使用不透明度和线条粗细来编码注意力连接的强度。我们还尝试可视化查询和键之间的所有权重，以更贴近地呈现 [53]，但这通常会产生过于拥挤、难以理解的结果。

6 系统实现

为了处理模型输入并计算注意力信息，我们使用 Hugging Face Transformers 库和 PyTorch。我们使用预训练的 BERT、GPT-2（small）和 ViT-16/32 实现，其模型权重来自 Google 和 OpenAI。对于每个自然语言处理数据集，我们随机采样 200 个句子（~10每个注意力头包含 k 个词元，包括查询和键）。由于图像注意力数据的计算规模较大，对于 ViT-32，我们每个注意力头显示 10 张图像（1000 个词元）；对于 ViT-16，则每个注意力头显示 4 张图像（1576 个词元）。可以将更大的数据集输入 AttentionViz，但这会增加系统延迟。目前，加载每个自然语言处理数据集需要 ~6 秒，加载 ViT-16 数据需要 ~10 秒。提取每个注意力头的查询和键向量嵌入后，我们生成相应的二维/三维 t-SNE、UMAP 和主成分分析坐标（第 5.1节）。为了为 ViT 中的图像块生成语义标签（例如“狗”或“背景”），我们使用 DeepLabv3 分割模型 [7]。总计，在 NVIDIA A100 GPU 上预处理每个 BERT/GPT-2 数据集需要 ~3小时；对于 ViT-32，则需要 ~30 分钟。

我们的最终 AttentionViz 原型由一个 Python/Flask 后端组成，该后端与使用 Vue 和 TypeScript 编写的前端进行通信。演示系统可访问：http://attentionviz.com。由于数据量庞大且受到浏览器内存限制的影响，我们通过后端使用 JSON 文件加载预先计算的注意力/投影信息。对于 ViT，后端还会执行图像处理（例如图块高亮和透明度调整），以便在前端显示。我们使用Deck.gl来可视化所得的查询-键联合嵌入。AttentionViz 具有高度可扩展性且与模型无关，允许用户向系统中添加新的 Transformer 和数据集。

7 研究发现与评估

我们通过三个应用场景以及领域专家的反馈，展示了 AttentionViz 的实用性。我们的场景旨在实现第4节中的目标，并展示 AttentionViz 如何为视觉和语言 Transformer 中的全局自注意力趋势提供洞见。

数据。对于 BERT/GPT-2，我们尝试了多个自然语言处理数据集，但针对应用场景重点选取了两个。我们使用 Wiki-Auto [23]作为基准，采样通用输入句子；使用 SuperGLUE AX_b [56] 探索文本蕴含任务特定的注意力模式。对于 ViT，我们从 ImageNet [44] 和 Microsoft COCO [28]，中采样图像，也使用合成图像数据。

用户访谈。我们邀请了 **E2** 和 **E3** 参加第二轮访谈，并邀请了两位新专家：**E6**（可解释性研究人员）和 **E7**（视觉科学博士生）。与第4节一样，我们分别对所有专家进行了单独访谈。首先，我们向专家快速演示了工具，并分享了我们的一些发现，请他们分享任何想法或见解（第 7.1-7.3节）。随后，我们请他们更概括地反馈 AttentionViz 的主要优势、劣势和创新之处（第 7.4节）。我们还询问了专家，对于将这一技术用于大规模可视化嵌入，可能有哪些扩展或应用。

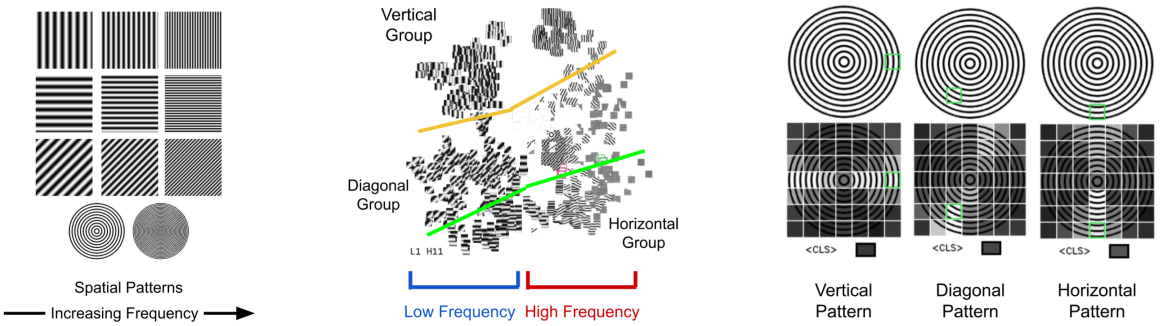


图 9：**左：** 包含不同频率和角度空间模式的数据集。**中：** 在单视图中，我们观察到，ViT-32 的一个注意力头会根据图像令牌所对应空间模式的频率和角度对它们进行排列。**右：** 图像视图中的注意力热图进一步证实了这些发现——角度相似的空间模式会彼此更加关注。

7.1 目标：理解机器视觉注意力

由于图像块数据具有天然的视觉属性，AttentionViz 对于发现视觉 Transformer 中有关注意力的新见解尤其有帮助 **[G1]**。

视觉注意力中的色调／亮度特化。 我们很好奇，是否有视觉注意力头专门处理基于颜色或基于亮度的模式。为验证这一点，我们向预训练的 ViT-32 模型输入合成颜色和亮度渐变图像（图 8），并将生成的查询和键令牌载入 AttentionViz。

在矩阵视图中浏览全局 PCA 模式时，我们发现了两个类似于彩色视觉和无彩色视觉的注意力头。一个注意力头似乎会根据亮度排列黑白图像令牌，另一个则会根据色调排列彩色图像块。我们的数据集包含各种方向的颜色和亮度渐变图像；我们发现，无论这些图像块在原始图像中的位置如何，相似的图像块都会聚集在联合嵌入空间中。图像视图中的注意力热图证实了这些发现：词元最关注颜色或亮度相同的其他词元。**E7** 对此很感兴趣，因为他此前研究过卷积神经网络（CNN）的颜色潜在空间，并表示有意使用我们的工具进一步探索 CNN 与 ViT 行为之间的差异。

频率滤波与角度检测。 频率和角度是图像数据的底层特征。为了探究视觉 Transformer 是否具有一个能根据这些特征关联视觉模式的注意力头，我们创建了频率和方向各异的正弦信号图像，并使用预训练的 ViT-32 模型对其进行处理。在矩阵视图中检查生成的查询和键嵌入后，我们发现了一个注意力头，它会根据图像令牌空间模式的频率（x 轴）和角度（y 轴）对图像令牌进行区分（图 9）。通过图像视图，我们观察到，同心圆图像中的词元会关注曲率相似的其他词元，这进一步证实该注意力头会根据角度关联视觉模式。

E7 认为这一结果很有意思，但鉴于我们关于色调/亮度的发现，这并不太令人意外；他们更好奇的是那些没有表现出这种“关注相似图像块”行为的注意力头。他们提出的一项实验是研究注意力调制，例如，如果同一个图像块（如垂直条纹）出现在两张图像的不同语境中（如斑马与雨伞），我们会看到独特的注意力模式吗？

模型层之间注意力距离的增加。 正如 [14], 所指出的，在视觉 Transformer 的较深层中，自注意力会在图像范围内关注更广泛的区域。我们通过 AttentionViz 中查询和键词元的交互式联合可视化验证了这一发现。借助矩阵视图（Matrix View），我们按照图像的“行”和“列”为图像块着色，从而在 ViT-32 的第 1 层和第 2 层中发现了四个注意力头，它们会将词元与其最近的空间邻居进行分组，即左侧、右侧、上方和下方的邻居。在第 3 层和第 4 层中，我们观察到了类似的位置注意力模式，但图像令牌关注的是同一行或同一列中的所有图像块，而不再局限于最近的邻居（图 10）。这表明，与使用方形滤波器处理图像的卷积神经网络不同，Transformer 中的自注意力机制通常按行处理图像，并逐列进行，类似于拉长的滤波器。

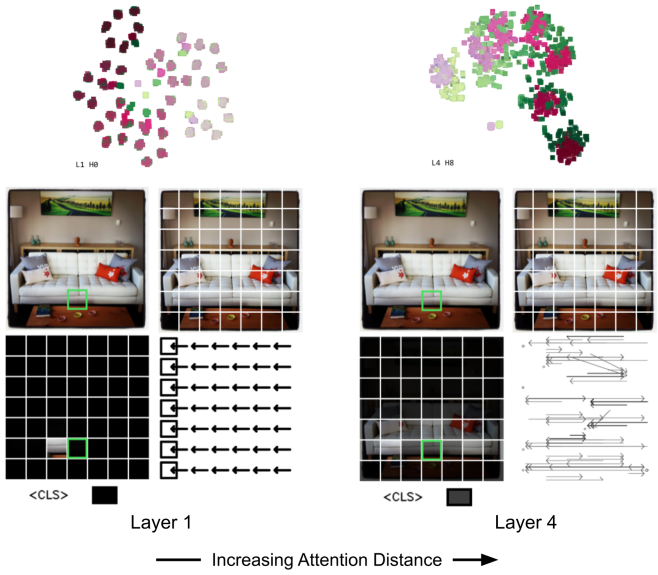


图10：按行给图像块着色，突显了 ViT-32 中的位置注意力模式。在**第1层**中，同一行且相邻列的词元形成小型簇。图像视图呈现出向左看的模式。在**第4层**中，词元根据所在行的位置形成大型簇。借助带箭头的连线，我们可以看到范围更广的双向注意力流。

7.2 目标：寻找全局注意力轨迹

为了理解语言Transformer中不同注意力头的自注意力模式如何变化 **[G2]**, , 我们使用 AttentionViz 探索了 BERT。

位置注意力特征。 通过 TSNE，我们观察到几个形状独特的注意力头，例如第3层中呈螺旋状的图（图5a）。例如，在单视图中，按归一化后的位置为第3层第9个注意力头着色，可以看出词元位置从螺旋外侧向内侧递增（图5b）。我们使用句子视图更仔细地检查了这一模式（图5c），确认这是一种基于位置的“下一词元”注意力模式。这种“螺旋”也反映了赋予 Transformer 的初始排序向量（第2节）。

随后，我们在矩阵视图中注意到其他可识别的“痕迹”，发现带有小“团簇”的图也编码了位置模式（图 11，左），并通过我们的离散位置着色验证了这一点。“螺旋”和“团簇”的区别似乎在于，词元是选择性地关注相隔一个位置的其他词元，还是关注相隔若干种不同位置的词元（图 5c）。同样，我们发现，在查询-键重叠度较高的注意力头中，词元通常会关注自身以及其他相同词元的实例，呈现出一种“关注自身”的模式。放大查看这些注意力头，我们可以看到相邻查询-键之间清晰的语义簇，如

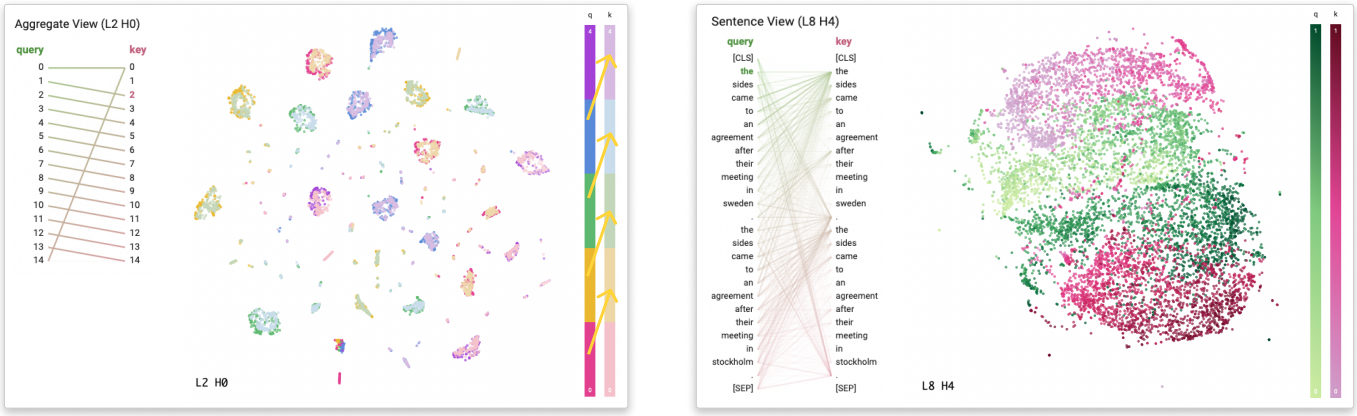


图 11：注意力的其他视觉痕迹。**左：**带有小型“簇”的注意力头通常比螺旋呈现出更紧密的位置模式。我们的离散位置编码根据每个词元的位置对 5 取模进行着色，突出了“下一个词元”的注意力趋势。**右：**查询和键的分层带状结构仅出现在 SuperGLUE AX_b 数据 [56], 中，表明模型对文本开头、结尾和中点具有很强的注意力。

图 6b 所示，这进一步支持了这一观察。

[29] 表明，较早的 Transformer 层包含有关线性词序的最多信息，这与我们的发现以及诸如 [9, 53] 之类的先前研究一致。在访谈过程中，**E2**、**E6** 和 **E7** 立即注意到了这些有趣的几何结构，尤其是螺旋，并且很想知道所观察到的结构在多大程度上纯粹源于位置因素。这启发专家提出了若干后续实验想法，例如操纵或移除 Transformer 模型中的位置嵌入，并观察我们的查询-键可视化如何变化。

任务特定痕迹。使用 AttentionViz 对多个数据集进行可视化后，我们发现，联合嵌入的形状在不同自然语言处理任务之间高度一致。然而，我们确实观察到一种视觉痕迹：它仅出现在 BERT 的某些较后层以及 SuperGLUE AX_b 数据中（图11，右）。点击其中一个这样的注意力头（第 8 层第 9 个注意力头）并按位置着色后，我们观察到了一个查询-键“夹层”：文本开头的键和查询堆叠在上方，随后是文本结尾处顺序颠倒的查询和键。

句子视图显示，文本的开头、中部和结尾获得了最多的注意力。整体图形形状和注意力模式表明，这些注意力头能够识别文本的“中点”，并区分不同句子，这与蕴含任务中的方式相似：比较两个句子，判断它们是否具有相近的含义。查询也主要关注同一句子中的键。[25,54] 显示，句法信息和任务特定信息在中后部的模型层中最为突出，这或许解释了该轨迹的独特性。

全局搜索模式。矩阵视图中的聚合搜索功能也可用于快速扫描和比较不同注意力头之间的注意力趋势 **[G2]**。我们发现，搜索结果中的模式反映了先前识别出的视觉注意力轨迹（图 6a）。例如，呈螺旋形或查询/键形成小簇的注意力头，其搜索结果更加分散，这表明它们具有底层的位置注意力模式。另一方面，具有“关注自身”注意力模式的注意力头只有一个搜索结果簇，凸显了同一词元的查询与键之间的强交互。

即使联合查询-键嵌入没有明显的形状特征，我们也发现，如果搜索结果簇只有少数几个，该注意力头可能会表现出更多语义行为；否则，很可能存在位置注意力模式。 [49] 指出，语义信息分布在 BERT 的各层中，这一点我们通过 AttentionViz 得到了证实。我们的所有专家都对这一工具的这一特性及其促进注意力模式比较的能力尤为赞赏。

7.3 目标：识别异常和意外行为

通过在 Attention-Viz 中探索联合查询-键嵌入，我们发现了一些异常模型行为 **[G3]**。

范数差异与空注意力。在 MatrixView 中探索 GPT-2 时，我们观察到在早期模型层中，某些查询和

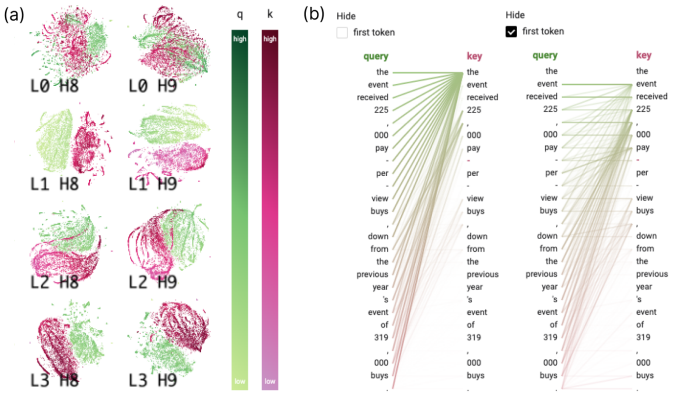
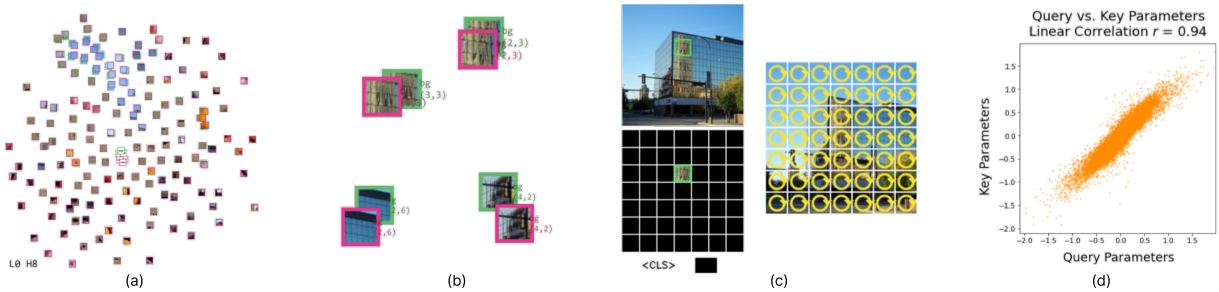


图 12：GPT-2 中的异常。**(a)** 在模型较早的层中，我们观察到许多注意力头的查询-键范数存在显著差异（例如，范数缩放前的*L1 H8*）。**(b)** 后续层中常见的“关注首个词元”模式示例。隐藏首个词元后，句子视图揭示了潜在的注意力行为。

即使经过键平移后（第 5.1.1 节），键簇仍然清晰分离。按照范数着色（测量值取自范数缩放步骤之前）后，我们发现，在许多注意力头中，查询向量和键向量的范数存在显著差异（图 12a）。当查询范数较小（即呈浅绿色）时，键范数往往较大（即呈深粉色），反之亦然。计算 GPT-2 与 BERT 中查询和键的平均范数差后，我们发现，前者的查询范数均值减去键范数均值为 **-4.59**（跨注意力头），而后者的平均差值仅为 **0.41**。我们的专家都无法解释这一发现：“查询和键的范数为何会有如此大的差异，确实说不通” **(E6)**。有趣的是，在我们观察到这一现象之后发表的一篇论文 [11] 指出，失控的查询和键范数是导致严重训练不稳定性的原因，这表明这一现象或许值得进一步研究。这一观察也启发了我们在第 5.1.1 节提出的缩放方法。

我们还注意到，在许多 GPT-2 注意力头中，大部分注意力都指向首个令牌（图12b），尤其是在较后的层中。 [54]简要提到，在 GPT-2 中，当注意力头所捕捉的语言属性没有出现在输入文本中时，首个令牌会被视为接收注意力的空位置。然而，这一现象仍缺乏充分探索，由此提出了另一个值得探讨的开放性可解释性问题。 **E2** 和 **E6** 都曾独立使用我们的工具注意到这一反常现象，而所有专家都对这一发现感到惊讶。 [55] 表明，在 Transformer 中剪除大多数注意力头，或许并不会显著影响模型性能，



(a) (b) (c) (d)

图13：在 ViT-32 中识别“关注自身”的注意力头。**(a)** 单视图显示，查询和键在联合嵌入空间中的分布较为稀疏。**(b)** 放大后可以看到，同一图像令牌的查询向量和键向量紧密重叠。**(c)** 图像视图显示，词元主要关注自身。**(d)** 比较查询投影层和键投影层学到的参数，证实它们学到了冗余投影。

这或许部分归因于这种占主导地位的空注意力模式。无论如何，AttentionViz 允许用户过滤掉指向首个令牌的注意力，从而揭示隐藏的查询-键交互。

“关注自身”的注意力头。 AttentionViz 还可以揭示视觉 Transformer 中出人意料的注意力模式。在矩阵视图中，我们在 ViT-32 的早期层发现了几个键-查询簇极为分散的注意力头（图 13a）。观察其中一个注意力头（第0层第8个注意力头）后，我们发现，同一词元的查询和键嵌入形成了一个小而密集的簇，其中每个查询-键对都与其他对清晰分离（图 13b）。从图像视图中的透明度热图可以看到，该图像块只关注自身（图 13c）。切换到带箭头的注意力连线后，我们发现这张图像的整体注意力模式是“关注自身”：该注意力头中图像令牌之间没有信息流动。

发现这种不规则的注意力模式后，我们通过相关性检验检查了查询矩阵和键矩阵的学习参数。我们发现两者的相似度很高（线性相关性 $=0.94$），这表明该 ViT 注意力头中的查询层和键层确实学到了冗余投影（图 13d）。**E3** 指出，这一发现有助于指导模型剪枝实验；AttentionViz 也可用于检测潜在的训练失败或其他异常。

7.4 用户反馈总结

矩阵视图的优点。几位专家认为，矩阵视图提供的“全局”视角是 AttentionViz 最具新意、也最有价值的部分。正如**E6** 所说：“它非常适合快速比较；想要一次可视化多个嵌入时，也不必再费心调整超参数。”**E7** 也提到，矩阵视图很有用，因为“对于规模较小的可视化，我可以自己写代码实现，但要处理更大的规模和更多数据就困难得多。”这些反馈表明，在其他机器学习场景中，大规模可视化和比较嵌入的思路也可能大有裨益。

联合查询-键嵌入的应用。 专家们提出了多种可视化技术的用例和扩展，表明其适用范围更广。例如，**E2** 建议可视化未训练或受损变换器中的模式，而**E3** 和 **E7**都希望可视化各自模型在训练过程中的注意力变化，这与我们的初始目标一致（第 4节）。 **E2** 和 **E6** 还建议调整我们的工具，以协助因果追踪，并解释说：“追踪整个模型中的注意力流，或许有助于进行假设检验。”同样，**E3** 表示有兴趣研究“不同注意力头中的两种注意力模式如何相互关联”，这当然可以用于可视化归纳头对。 **E2** 指出，添加一种“量化两个注意力头之间相似度”的方法可能会有所帮助，而 **E6** 则建议为了模型剪枝而“测量或可视化注意力头中的随机性”。

嵌入投影——究竟该不该信任？ E3 强调了使用投影方法所面临的挑战。尽管他们赞赏我们发现的醒目几何图案（例如螺旋），**E3** 仍对如何解读这些可视化结果表示怀疑，因为 t-SNE 和 UMAP 等技术会造成失真：“我怎么知道自己看到的是否可信？”这凸显了将可视化洞见与可执行的干预措施联系起来的重要性；或许可以扩展我们的工具，使其除了支持探索之外，还能支持假设检验。

灵活性与易用性的权衡。E2 指出，AttentionViz“用起来很顺手，也很容易自定义”，与“现有可视化工具学习和使用起来都过于复杂”形成对比。然而，一些专家，例如**E6**，仍然担心“展示所有特征和注意力头可能会让人不知所措……有没有办法概括这些信息？或者更聚焦于某项特定任务？”**E7** 补充道：“我想知道有没有更快捷、更易理解的方式来标注注意力头”，并建议采用一种更接近特征可视化的方案 [36]。我们将 AttentionViz 设计为一种灵活的工具（例如，支持在不同 Transformer 中进行注意力分析，并可采用不同粒度），但看来，我们设计中的灵活性与易用性权衡 [27] 仍有改进空间。

其他交互模式。 一些专家建议增加交互模式，例如即时推理 **(E3)**，或对圈出的查询和键簇进行进一步降维，以揭示更多信息并开展细粒度分析 **(E2)**。 **E7** 强调了允许用户直接向系统上传新数据集的重要性：“这个工具可以更加强大……人们会希望通过添加自己的图像等方式进行更多探索。”

8 结论与未来工作

在本研究中，我们提出了一种基于查询和键的联合嵌入空间、用于可视化 Transformer 自注意力的新技术。借助这项技术，我们构建了 AttentionViz（演示：http://attentionviz.com），这是一款交互式可视化工具，并用它深入探究语言和视觉 Transformer 中的注意力。例如，我们发现了 ViT 中新颖的色调／频率行为，以及 GPT-2 中显著的查询与键范数差异。尽管我们的方法专为自注意力设计，但经过适当调整向量归一化方式后，也可推广至其他注意力机制（例如交叉注意力）。同样，我们着重研究语义模式，但 AttentionViz 也可用于研究句法特征，例如利用自然语言处理库添加词性等元数据。

专家反馈还指出了未来工作的几个方向，例如探索管理多个嵌入可视化复杂性的方法，并引导用户关注感兴趣的特征。我们的系统目前受限于数据预计算时间和内存需求，但我们计划提升 AttentionViz 的可扩展性，并允许用户动态添加新的输入。规模足够大的随机样本（例如几千个词元）足以呈现大规模的全局现象，但更大的数据集和模型可能会揭示更多语义洞见，正如 [38] 所提出的那样。尝试其他数据采样方法也可能有所帮助（例如，减少齐夫定律带来的偏差 [62]）。

未来研究的另一个自然方向，是探索如何纳入每个注意力头中的值向量信息 [52]。这些值向量是注意力机制的重要组成部分，但目前尚不清楚如何结合查询和键来对其进行可视化。找到合适的可视化方法，或许能让我们更深入地了解注意力头的工作方式。最后，尽管 AttentionViz 是一种探索性工具，但将其改造用于假设检验和／或因果追踪，或许能为实际的模型调试提供支持。

致谢

我们感谢所有参与用户访谈的人士，感谢他们付出的时间并分享了宝贵见解。我们也感谢匿名审稿人，他们提出了详尽且富有建设性的意见，帮助我们改进本文。最后，我们感谢哈佛 Insight + Interaction Lab 的成员在整个项目期间给予的深思熟虑的反馈与支持。

参考文献

[1] E. Aflalo, M. Du, S.-Y. Tseng, Y. Liu, C. Wu, N. Duan, and V. Lal. Vl-interpret：用于解读视觉语言变换器的交互式可视化工具。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 21406–21415 页。IEEE，新奥尔良，2022。doi: 10.1109/CVPR52688.2022.02072 2[2] D. L. Arendt, N. Nur, Z. Huang, G. Fair, and W. Dou. Parallel embeddings：一种用于对比学习到的表征的可视化技术。载于第 25 届智能用户界面国际会议论文集，第 259–274 页。ACM，卡利亚里，2020。doi: 10.1145/3377325. 3377514 3[3] A. Boggust, B. Carter, and A. Satyanarayan. 嵌入比较器：通过小多图可视化全局结构和局部邻域的差异。载于第 27 届智能用户界面国际会议论文集，第 746–766 页。ACM，赫尔辛基，2022。doi: 10.1145/3490099.[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. 自监督视觉变换器中涌现的特性。载于《IEEE/CVF国际计算机视觉会议论文集》，第 9650–9660 页。IEEE，蒙特利尔，2021。doi: 10.1109/ICCV48922. 2021.00951 2[5] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah. Activation Atlas. Distill,

4 3 2019 10 23915 00015 1 ( ), . doi: . /distill.

[6] H. Chefer, S. Gur, and L. Wolf. 超越注意力可视化的 Transformer 可解释性。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 782–791 页。IEEE，线上，2021。doi: 10.1109/CVPR46437.2021.00084 3[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. 重新思考用于语义图像分割的空洞卷积。arXiv 预印本，2017。doi: 10.48550/arXiv.1706.05587 6 [8] E. A. Chi, J. Hewitt, and C. D. Manning. 在多语言BERT中寻找普适语法关系。载于计算语言学协会第 58 届年会论文集，第 5564–5577 页。ACL，线上，2020。doi: 10.18653/v1/2020.acl-main.493 2, 3[9] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. BERT 关注什么？对 BERT 注意力的分析。载于《2019 年 ACL BlackboxNLP研讨会：自然语言处理神经网络分析与解读》论文集，第 276–286 页。ACL，佛罗伦萨，2019。doi: 10.18653/v1/W19-4828 2, 8[10] J.-B. Cordonnier, A. Loukas, and M. Jaggi. 自注意力与卷积层之间的关系。载于第八届国际学习表征会议。亚的斯亚贝巴，2020。doi: 10.48550/arXiv. 1911.03584 2[11] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, et al. 将视觉变换器扩展至 220 亿参数。arXiv 预印本，2023。doi: 10. 48550/arXiv.2302.05442 1, 8[12] J. F. DeRose, J. Wang, and M. Berger. Attention Flows：分析和比较语言模型中的注意力机制。《IEEE可视化与计算机图形学汇刊》，27(2):1160–1170, 2021。doi: 10 .1109/TVCG.2020.3028976 2[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT：用于语言理解的深度双向变换器预训练。arXiv 预印本，2018 10 48550 1810 04805 1, 2 。doi: . /arXiv. . [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Un- terthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. 一张图像胜过 16x16 个词：大规模图像识别变换器。载于第九届国际学习表征会议。线上，

2021 10 48550 2010 11929 1, 2, 7 。doi: . /arXiv. . N.

[15] Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, et al. 叠加玩具模型。arXiv 预印本，

2022 10 48550 2209 10652 2 。

[16] doi: . /arXiv. . N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al. 一个数学框架——适用于 Transformer 电路。Transformer Circuits Thread，2021。https: // transformer-circuits.pub/

2021 2 /

[17] framework/index.html。A. Ghiasi、H. Kazemi、E. Borgnia、S. Reich、M. Shu、M. Goldblum、A. G. Wilson 和 T. Goldstein。视觉 Transformer 学到了什么？一次视觉探索。arXiv 预印本，

2022 10 48550 2212 06727 3 。doi: . /

[18] arXiv. . F. Hohman、M. Kahng、R. Pienta 和 D. H. Chau。深度学习中的可视分析：面向未来前沿的探询式综述。IEEE可视化与计算机图形学汇刊，25(8):2674–2693, 2018。doi: 10 .1109/TVCG.2018.2843369 3[19] F. Hohman、H. Park、C. Robinson 和 D. H. P. Chau。Summit：通过可视化激活和归因摘要来扩展深度学习可解释性。IEEE可视化与计算机图形学汇刊，

26 1 1096 1106, 2019 10 1109 2019 2934659 3 ( ): – 。doi: . /TVCG. .

[20] B. Hoover、H. Strobelt 和 S. Gehrmann。exBERT：用于探索 Transformer 模型中学习到的表征的可视分析工具。载于《第 58 届计算语言学协会年会论文集：系统演示》，第 187–196 页。ACL，线上，2020。doi: 10. 18653/v1/2020.acl-demos.22 1, 2[21] T. Jaunet、C. Kervadec、R. Vuillemot、G. Antipov、M. Baccouche 和 C. Wolf。Visqa：对 Transformer 中的视觉与语言推理进行 X 射线分析。IEEE可视化与计算机图形学汇刊，

28 1 976 986, 2022 10 1109 2021 3114683 2 ( ): – 。doi: . /TVCG. .

[22] X. Ji、Y. Tu、W. He、J. Wang、H.-W. Shen 和 P.-Y. Yen。Usevis：信息检索中基于注意力的神经嵌入的可视分析。Visual Informatics，

5 2 1 12, 2021 10 1016 2021 03 003 2 ( ): – 。doi: . /

[23] j.visinf. . C. Jiang、M. Maddela、W. Lan、Y. Zhong 和 W. Xu。用于文本简化中句子对齐的神经 CRF 模型。载于《第 58 届计算语言学协会年会论文集》，第 7943–7960 页。ACL，线上，2020。doi: 10.18653/v1/2020.acl-main.709 1,[24] I. T. Jolliffe。回归分析中的主成分，第 129–155 页。Springer，纽约，

1986 10 1007 978 1 4757 1904 8_8 4 。doi: . / - - - - O.

[25] Kovaleva、A. Romanov、A. Rogers 和 A. Rumshisky。揭示 BERT 的黑暗秘密。载于《2019 年自然语言处理经验方法会议暨第九届自然语言处理国际联合会议论文集》，第 4365–4374 页。ACL，中国香港，

2019 10 18653 1 19 1445 2, 8 。doi: . /v /D - Q. Li、K. S.

[26] Njotoprawiro、H. Haleem、Q. Chen、C. Yi 和 X. Ma。Embeddingvis：一种用于比较网络嵌入检查的可视分析方法。载于《2018 IEEE 可视分析科学与技术会议》，第 48–59 页。IEEE，柏林，2018。doi: 10.1109/VAST.2018. 8802454 3[27] W. Lidwell、K. Holden 和 J. Butler。《设计通用原则》修订更新版：125 种增强可用性、影响感知、提升吸引力、做出更佳设计决策并通过设计进行教学的方法。Rockport Pub，贝弗利，2010 9 。T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Dollár 和 C. L. Zitnick。Microsoft COCO：上下文中的常见物体。载于《欧洲计算机视觉会议》，第 740–755 页。Springer，苏黎世，[28]

2014 10 1007 978 3 319 10602 1_48 4 6 。doi: . / - - - - , Y. Lin、Y. C. Tan

[29] 和 R. Frank。芝麻开门：深入了解 BERT 的语言知识。载于《2019 ACL Black-boxNLP 研讨会论文集：分析与解释 NLP 神经网络》，第 241–253 页。ACL，佛罗伦萨，2019。doi: 10.18653/v1/W19-4825 8[30] S. Liu、T. Li、Z. Li、V. Srikumar、V. Pascucci 和 P.-T. Bremer。针对自然语言推理和机器阅读理解中基于注意力的模型的视觉探询。载于《2018 年自然语言处理经验方法会议论文集：系统演示》，第 36 41 2018 10 18653 1 18 2007 1, 2 – 页。ACL，布鲁塞尔，[31] 。doi: . /v /D - J. Ma、Y. Bai、B. Zhong、W. Zhang、T. Yao 和 T. Mei。视觉化并理解视觉 Transformer 中的图像块交互。IEEE Transactions on Neural Networks and Learning Systems，第 1–10 页，2023。doi: 10. 1109/TNNLS.2023.3270479 2[32] L. McInnes、J. Healy 和 J. Melville。UMAP：用于降维的均匀流形近似与投影。arXiv 预印本，2018。doi: 10.48550/arXiv.1802.03426 4[33] P. Mehrani 和 J. K. Tsotsos。视觉 Transformer 中的自注意力执行的是知觉分组，而非注意力。arXiv 预印本，2023。doi: 10. 48550/arXiv.2303.01542 2[34] A. Miaschi、D. Brunato、F. Dell'Orletta 和 G. Venturi。神经语言模型的语言学剖析。载于《第 28 届国际计算语言学会议论文集》，第 745–756 页。国际计算语言学委员会，巴塞罗那，2020。doi: 10.

18653/v1/2020.coling-main.65 2
[35] M. Khan、M. Naseer、M. Hayat、R. Anwer、F. Shahbaz、S. Khan 和 M.-H. Yang。《视觉 Transformer 的引人注目的特性》。载于《神经信息处理系统进展》第34卷，第23296–23308页。Curran Associates, Inc.，线上，2021年。doi: 10.48550/arXiv.2105.10497 2[36] C. Olah、A. Mordvintsev 和 L. Schubert。《特征可视化》。Distill, 2 11 2017 10 23915 00007 9 ( ), 。doi: 。/[37] distill。C. Olsson、N. Elhage、N. Nanda、N. Joseph、N. DasSarma、T. Henighan、B. Mann、A. Askell、Y. Bai、A. Chen 等。《上下文学习与归纳注意力头》。arXiv 预印本，2022年。doi: 10.48550/arXiv.2209.11895 2[38] M. Oquab、T. Darcet、T. Moutakanni、H. Vo、M. Szafraniec、V. Khalidov、P. Fernandez、D. Haziza、F. Massa、A. El-Nouby 等。《DinoV2：无需监督学习鲁棒视觉特征》。arXiv 预印本，2023年。doi: 10. 48550/arXiv.2304.07193 9[39] C. Park、I. Na、Y. Jo、S. Shin、J. Yoo、B. C. Kwon、J. Zhao、H. Noh、Y. Lee 和 J. Choo。《Sanvis：用于理解自注意力网络的可视分析》。载于《2019年 IEEE 可视化会议》，第146–150页。IEEE，温哥华，2019年。doi: 10.1109/VISUAL.2019.8933677 2, 6[40] N. Park 和 S. Kim。《视觉 Transformer 如何工作？》。载于《第十届学习表征国际会议》。线上，2022年。doi: 10. 48550/arXiv.2202.06709 2[41] A. Radford、K. Narasimhan、T. Salimans、I. Sutskever 等。《通过生成式预训练提升语言理解能力》。OpenAI 博客，2018年。https://openai.com/research/language-unsupervised。 2
[42] A. Radford、J. Wu、R. Child、D. Luan、D. Amodei、I. Sutskever 等。《语言模型是无监督的多任务学习者》。OpenAI 博客，2019年。https://openai.com/research/better-language-models。 1, 2
[43] E. Reif、A. Yuan、M. Wattenberg、F. B. Viegas、A. Coenen、A. Pearce 和 B. Kim。《可视化并测量 BERT 的几何结构》。载于《神经信息处理系统进展》第32卷。Curran Associates, Inc.，温哥华，2019 10 5555 3454287 3455058 2, 3 。[44] doi: 。/。O. Russakovsky、J. Deng、H. Su、J. Krause、S. Satheesh、S. Ma、Z. Huang、A. Karpathy、A. Khosla、M. Bernstein 等。《ImageNet 大规模视觉识别挑战赛》。International Journal of Computer Vision，115：211–252, 2015 10 1007 11263 015 0816 6 。doi: 。/s - - -y R. Sevastjanova、[45] E. Cakmak、S. Ravfogel、R. Cotterell 和 M. El-Assady。《语言模型适配的可视化比较》。IEEE可视化与计算机图形学汇刊，29(1)：1178–1188，2022年。doi: 10.1109/TVCG.2022.3209458 3[46] V. Sivaraman、Y. Wu 和 A. Perer。《Emblaze：通过交互式比较嵌入空间阐明机器学习表征》。载于《第27届智能用户界面国际会议》，第 418 432 2022 10 1145 3490099 3511137 3 –页。ACM，赫[47] 尔辛基，。doi: 。/。D. Smilkov、N. Thorat、C. Nicholson、E. Reif、F. B. Viégas 和 M. Wattenberg。《嵌入投影器：嵌入的交互式可视化与解读》。arXiv 预印本，2016年。doi: 10.48550/arXiv.1611.05469 3, 5[48] H. Strobelt、S. Gehrmann、M. Behrisch、A. Perer、H. Pfister 和 A. M. Rush。《Seq2seq-vis：用于序列到序列模型的可视化调试工具》。IEEE可视化与计算机图形学汇刊，25 1 353 363, 2019 10 1109 2018 2865044 2 ( ): –。doi: 。/TVCG.。[49] I. Tenney、D. Das 和 E. Pavlick。《BERT 重新发现了经典自然语言处理流程》。载于《第57届计算语言学协会年会论文集》，第4593–4601页。ACL，佛罗伦萨，2019年。doi: 10.18653/v1/P19-1452 2, 8[50] L. van der Maaten 和 G. Hinton。《使用 t-SNE 可视化数据》。Journal of Machine Learning Research，9(86)：2579–2605，2008年。http://jmlr.org/papers/v 9 08 4 /vandermaaten a.html。A.[51] Vaswani、S. Bengio、E. Brevdo、F. Chollet、A. N. Gomez、S. Gouws、L. J ones、Ł. Kaiser、N. Kalchbrenner、N. Parmar 等。《用于神经机器翻译的 Tensor2tensor》。arXiv 预印本 arXiv: *1803 07416*, 2018 1, 2 。。A.[52] Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. u. Kaiser 和 I. Polosukhin。《注意力就是你所需要的一切》。载于《神经信息处理系统进展》第30卷。Curran Associates, Inc.，长滩，2017 10 48550 1706 03762 1, 2, 9 。doi: 。/arXiv.。J.[53] Vig。《Transformer 模型中注意力的多尺度可视化》。载于《第57届计算语言学协会年会论文集：系统演示》，第37–42页。ACL，佛罗伦萨，2019年。doi: 10.18653/v1/P19-3007 1, 2, 5, 6, 8[54] J. Vig 和 Y. Belinkov。《分析 Transformer 语言模型中的注意力结构》。载于《2019年 ACL BlackboxNLP研讨会：自然语言处理神经网络的分析与解读论文集》，第63–76页。ACL，佛罗伦萨，2019 10 18653 1 19 4808 8 。doi: . /v /[55] W - E. Voita、D. Talbot、F. Moiseev、R. Sennrich 和 I. Titov。《分析多头自注意力：专门的注意力头承担了主要工作，其余部分可以剪除》。载于《第57届计算语言学协会年会论文集》，第5797–5808页。ACL，佛罗伦萨，2019。doi: 10. 18653/v1/P19-1580 8[56] A. Wang、Y. Pruksachatkun、N. Nangia、A. Singh、J. Michael、F. Hill、O. Levy 和 S. R. Bowman。《SuperGLUE：一个更具挑战性的通用语言理解系统基准测试》。载于《神经信息处理系统进展》，第32卷，第3266–3280页。Curran Associates Inc.，温哥华，2019。doi: 10.5555/3454287.3454581 6, 8[57] K. Wang、A. Variengien、A. Conmy、B. Shlegeris 和 J. Steinhardt。《真实场景中的可解释性：GPT-2 small 中用于间接宾语识别的电路》。arXiv 预印本，2022 10 48550 2211 00593 2 。doi: . /[58] arXiv. . Y. Wang、S. Liu、N. Afzal、M. Rastegar-Mojarad、L. Wang、F. Shen、P. Kingsbury 和 H. Liu。《生物医学自然语言处理中词嵌入的比较》。生物医学信息学杂志，87 12 20, 2018 10 1016 2018 09 008 3 : –。doi: . /j.jbi. . . Z. J. Wang 、 R. Turko 和 D. H. Chau。《 Dodrio：通过交互式可视化探索 Transformer 模型》。载于《第59届计算语言学协会年会暨第11届自然语言处理国际联合会议：系统演示》，第132–141页。ACL，线上，2021。doi: 10.18653/v1/2021 .acl-demo.16 2[60] J. Wei、Y. Tay、R. Bommasani、C. Raffel、B. Zoph、S. Borgeaud、D. Yo- gatama、M. Bosma、D. Zhou、D. Metzler、E. H. Chi、T. Hashimoto、O. Vinyals、P. Liang、J. Dean 和 W. Fedus。《大型语言模型的涌现能力》。《机器学习研究汇刊》，2022。doi: 10.48550/arXiv.2206.07682 1[61] S. Yang、Z. Quan、M. Nie 和 W. Yang。《Transpose：通过 Transformer 实现关键点定位》。载于《IEEE/CVF国际计算机视觉会议论文集》，第11802–11812页。蒙特利尔，2021。doi: 10.1109/ ICCV48922.2021.01159 3[62] G. K. Zipf。《语言中相对频率原理的精选研究》。哈佛大学出版社，剑桥，1932。doi: 10.4159/harvard.9780674434929 9