更多AI工具可直接访问：https://www.faxianai.com/

中学生能看懂：快手「可灵」和「Sora」背后 DiT 技术

⚽ 金色传说大聪明 赛博禅心 *2024-06-09*

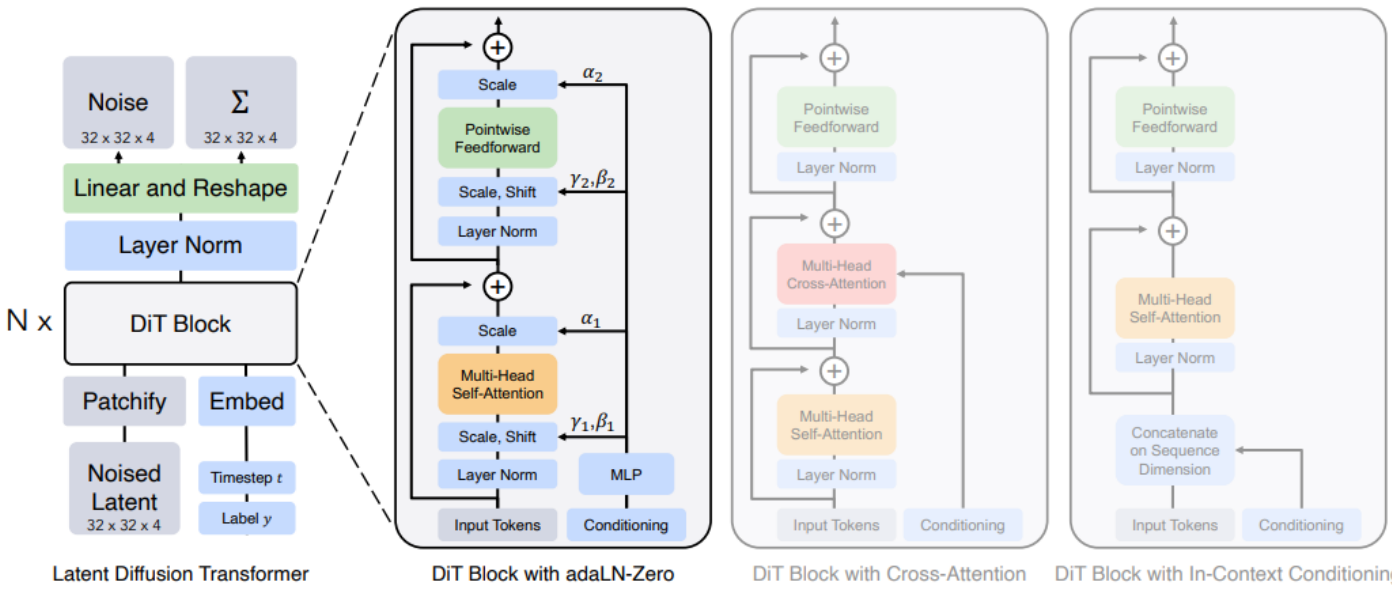
原文链接：https://mp.weixin.qq.com/s/2PrMgNAL0Er_vNjhqSbTHQ

可灵AI访问入口：https://www.faxianai.com/ai/1098.html

写在前面

本文作者：大聪明 GPT v2.0

下为正文，文末附 prompt



Diffusion 模型的运作，像是在玩一个“加噪声再还原”的游戏。想象一下，我们把一张清晰的图片比作一杯清水。Diffusion 模型一开始会慢慢往这杯清水里滴入墨汁，让水变得越来越浑浊，最终变成一杯完全看不清的墨水。这个过程就是“加噪声”。

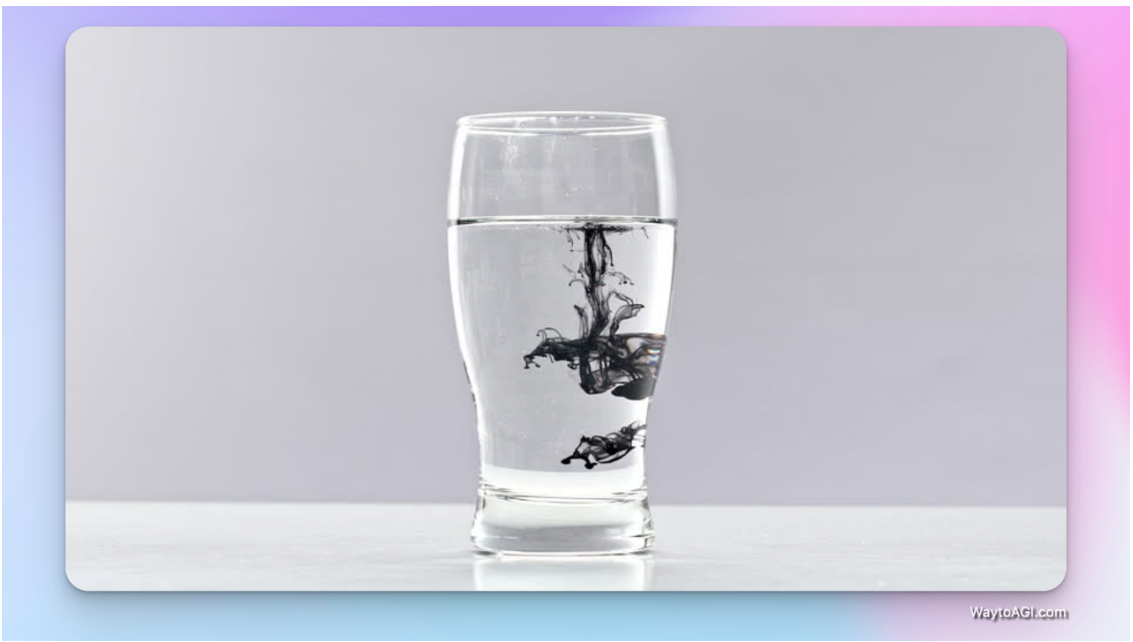
有趣的是，Diffusion 模型接下来要做的，就是想办法把这杯浑浊的墨水还原成最初的清水。怎么做呢？它会学习“加噪声”的逆过程，一步一步地去除噪声，就像把墨水从水中慢慢分离出来一样，最终恢复成清晰的图片。

为了更好地学习噪声和图像之间的关系，DiT 模型巧妙地引入了 Transformer 架构。Transformer 最厉害的地方在于它拥有“注意力机制”，就像我们在阅读文章时，会特别关注一些关键词一样，注意力机制能让模型学会抓住数据中的关键信息。

DiT 模型将 Transformer 架构应用于 Diffusion 模型中，就好像给模型戴上了一副“眼镜”，让它能够更清楚地看到噪声和图像之间的联系，从而更精准地去除噪声，生成更高质量的图像。

Diffusion 模型的基本原理

Diffusion 模型的精妙之处在于它颠覆了传统的绘画方式。想象一下，我们不是在一张白纸上作画，而是先将一幅完整的画作逐渐添加噪声，就像将颜料一点一点地涂抹在画布上，最终让画面变得一片模糊，完全看不出原本的模样。这个过程就像墨水在清水中扩散，最终均匀地分散，再也无法分辨最初的墨滴。



更令人惊叹的是，Diffusion 模型不仅学会了如何“破坏”，更掌握了如何“重建”。它能够将被噪声污染的模糊图像一步步地还原，最终恢复成最初清晰的画面。这就好比将混着墨水的清水一点点净化，Diffusion 模型会仔细分析图像中颜色的分布规律，判断哪些颜色是后期添加的噪声，然后像魔法师一样，将这些噪声逐一擦除。

为了更精细地控制“添加噪声”和“去除噪声”的过程，Diffusion 模型将这两个过程都分解成许多个小步骤。在每一个步骤中，添加或去除的噪声都非常少，就像画家用画笔轻轻地描绘，一点一点地改变着画面。模型会记录下每一步操作后的图像状态，我们称之为“时间步长”。“时间步长”就像是指引模型还原图像的路标，告诉模型当前处于还原过程的哪个阶段。模型只需要根据“时间步长”的

信息，就可以知道下一步应该去除哪些噪声，最终像倒放电影一样，将被噪声污染的图像逐步还原成最初的清晰模样。

Transformer 架构的引入，为 Diffusion 模型的图像生成能力带来了质的飞跃。

Transformer 架构的作用

前面我们了解到 Diffusion 模型的工作原理，就像是一位技艺高超的画家，先是用噪音把图像完全遮盖，再一步步地去除噪音，最终创作出一幅完整的作品。那么，在 DiT 模型中，是谁充当着这位画家的角色呢？答案就是 Transformer。

Transformer 模型就像是由许多个“注意力模块”搭建起来的，每个模块都像是一个聪明的观察者，能够分析图像中不同部分之间的关系。还是以“我喜欢吃苹果”这句话为例，Transformer 会注意到“喜欢”和“苹果”之间的联系，从而更加重视“苹果”这个词，而相对弱化“吃”这个词。

在 DiT 图像生成模型中，Transformer 就如同一位经验丰富的指挥家，它引导模型关注图像的关键信息。例如，天空的颜色和云朵的形状之间存在着某种联系，Transformer 能够学习到这种联系，并指导模型生成更加真实自然的图像。

DiT 的训练过程

Transformer 架构赋予了 DiT 模型强大的图像处理能力。那么，我们是如何训练 DiT 模型的呢？

简单来说，训练过程就像教一位绘画新手学习创作。首先，我们会向 DiT 模型展示大量的清晰图像，就像给新手展示大师级的作品，让模型明白学习的目标。接着，我们逐步向图像添加噪声，就好像在画布上逐渐涂抹颜料，最终图像会被噪声完全覆盖，变得难以辨认。

DiT 模型的任务就是从这些被噪声遮蔽的图像中，还原出原始的清晰图像。这个过程就像新手画家临摹大师作品，从模糊的轮廓开始，逐步完善细节和色彩，最终完成作品。为了记录每一步添加噪声的程

度，我们引入了“时间步长嵌入”的概念，就像记录学习过程的日记。模型可以根据“时间步长”信息，逐步去除图像噪声，最终还原出清晰的原始图像。

当然，这个训练过程需要海量的数据和强大的计算资源，就像绘画需要丰富的素材和宽敞的工作室。模型通过“观摩”海量的图像素材，学习如何表达物体、颜色和纹理。强大的计算资源则为模型提供了高效的学习环境，使其能够快速处理数据，不断优化算法，最终成为技艺精湛的“绘画大师”。DiT 模型正是通过这样的训练过程，掌握了从噪声中生成清晰图像的能力。

利用学习到的强大能力，DiT 模型在多个图像生成任务中表现出色。

DiT 的应用

不仅训练过程高效，DiT 在实际应用中也展现出强大的图像生成能力。只需输入简单的文字描述，DiT 就能生成逼真的图像，无论是壮丽的自然景观，还是栩栩如生的人物肖像，甚至是各种可爱的动物，DiT 都能轻松应对。

例如，用户可以输入“一只毛茸茸的棕色小猫，有着明亮的蓝色眼睛”，DiT 就能根据描述生成一张符合要求的小猫图片，甚至可以细致到毛发的纹理和眼神的光彩。DiT 就像一个技艺精湛的画家，能够将文字转化为充满细节的图像，令人惊叹。

除了生成全新的图像，DiT 在图像修复方面也表现出色。对于一些珍贵的旧照片，由于时间久远，难免会出现划痕、污渍甚至破损，DiT 可以智能识别图像中的瑕疵，并进行修复，使其恢复到原本的清晰状态。

例如，一张老照片上的人物面部被污渍遮挡，DiT 可以分析照片中其他部分的人物特征，并根据这些特征“脑补”出被遮挡的部分，最终修复出完整清晰的人物面部，让珍贵的记忆重现光彩。

不仅如此，DiT 还可以赋予图像不同的艺术风格。例如，将一张普通的照片转换成梵高风格的油画，充满艺术气息；也可以将一张白天拍摄的照片转换成夜晚的效果，呈现出截然不同的氛围。DiT 为图像创作和编辑提供了全新的可能性，让用户可以尽情发挥想象力，创造出独一无二的图像作品。

DiT 的优势

DiT 不仅能生成更美的图片，还能生成更多样的图片。比如，我们可以给 DiT 输入一段文字，让它根据文字描述生成相应的图片；也可以给它一张简单的草图，它就能补充细节，生成完整的图片。

传统的 Diffusion 模型在处理图像细节时常常力不从心，生成的图像不够逼真。而 DiT 利用 Transformer 强大的全局建模能力，能捕捉到图像中更细微的特征，例如人物的头发丝、动物的毛发等等，这些细节在 DiT 生成的图像中都能清晰地展现出来。

除此之外，DiT 生成图像的速度也比传统的 Diffusion 模型更快。这是因为 Transformer 架构可以并行处理图像数据，大大提高了图像生成的效率。生成一张高清图片，传统的 Diffusion 模型可能需要几分钟甚至更长时间，而 DiT 只需要短短几秒钟。

DiT 的性能评估

DiT 模型拥有一项重要优势：它能生成十分逼真的图像。如何判断图像的逼真程度呢？我们可以借助一些量化指标，例如 FID（Fréchet Inception Distance）。FID 分数越低，代表生成图像与真实图像越相似，图像质量也就越高。

我们可以把 FID 的工作原理想象成这样：假设我们有一位经验丰富的水果鉴赏家，他能够轻易分辨出真实苹果和图片的区别。现在，我们用 DiT 模型生成一些苹果图片，并把它们混入真实的苹果中。如果这位鉴赏家无法分辨出哪些是真实苹果，哪些是 DiT 生成的图片，那就说明 DiT 模型生成的图片非常逼真。

当然，现实中我们不可能每次都请来一位鉴赏家来评判图像的质量。因此，我们会借助预先训练好的图像分类模型（例如 Inception 网络）来充当“鉴赏家”的角色。这个模型就像经验丰富的鉴赏家一样，能够识别图像的各种特征，并对其进行分类。

具体来说，我们会将真实图像和 DiT 生成的图像分别输入这个模型，并提取模型最后一层的特征向量。这些特征向量就像是对图像进行高度概括后的描述，包含了图像的关键特征信息。FID 分数则用于

衡量真实图像和生成图像的特征向量之间的距离。距离越近，说明 DiT 模型生成的图像越逼真。

大量实验数据表明，DiT 模型在 FID 指标上的表现非常出色，生成的图像质量甚至可以超越其他先进的图像生成模型。这说明 DiT 模型能够很好地捕捉图像的特征，并生成以假乱真的图像。

基于 DiT 模型在图像生成质量方面的优异表现，我们可以预见其在未来拥有广阔的应用前景。

DiT 技术的未来发展

DiT 模型在图像生成质量和分辨率上还有很大提升空间，关键在于学习更复杂的图像特征。未来的 DiT 模型将会更强大，生成的图像也会更逼真、更有创意，甚至可以与照片相媲美。科学家们正在尝试将 DiT 与其他人工智能技术结合，例如强化学习和对抗生成网络 （GAN）。这些结合有可能进一步提高 DiT 的性能，开拓新的应用领域，例如生成更复杂的场景和图像。

DiT 技术在游戏开发、虚拟现实和增强现实等领域拥有巨大的应用潜力。想象一下，游戏开发者可以使用 DiT 生成逼真的游戏场景和角色，为玩家带来更身临其境的游戏体验。在虚拟现实和增强现实领域，DiT 可以创建与现实世界无缝融合的虚拟对象和环境，为用户带来更具沉浸感和互动性的体验。

当然，DiT 技术的进一步发展也面临着一些挑战。例如，目前的 DiT 模型计算复杂度较高，需要大量的计算资源才能运行，这限制了其在一些资源受限设备上的应用。因此，降低 DiT 模型的计算复杂度是一个重要的研究方向。此外，确保 DiT 生成的图像符合伦理和道德标准也是一个需要关注的问题。

DiT 技术的挑战和局限

DiT 模型虽然潜力巨大，但其发展也面临着一些挑战。

训练一个 DiT 模型需要强大的计算资源和海量的数据。特别是对于那些性能卓越的 DiT 模型，往往需要依赖价格昂贵的硬件设备，训练过程也需要耗费大量的时间。这对许多研究者和开发者来说，都是一个不小的门槛，限制了 DiT 技术的普及和应用。

此外，DiT 模型的内部机制非常复杂，我们很难理解它究竟是如何做出决策，如何生成图像的。这种“黑箱”特性使得我们难以解释模型背后的逻辑，也为模型的进一步优化和应用带来了挑战。

DiT 模型强大的图像生成能力也引发了人们的担忧。它可以被用来生成以假乱真的虚假信息，比如 Deepfake，从而误导公众，甚至引发社会问题。这就需要我们开发相应的技术手段来鉴别和防范此类风险，以确保 DiT 技术的合理和负责任地使用。

虽然 DiT 模型生成的图像质量已经取得了很大的进步，但仍有提升的空间。有时生成的图像仍然会出现细节模糊、不够逼真等问题。

总结与展望

DiT 模型在图像生成领域取得了显著的进展，生成的图像质量非常高，可以应用于图像编辑、风格转化等多个方面。传统的图像生成方法难以学习复杂的图像特征，导致生成的图像质量不够理想。DiT 模型将 Transformer 架构引入 Diffusion 模型，成功突破了传统方法的瓶颈，能够学习更复杂的图像特征，从而生成更逼真、更富有细节的图像。

DiT 模型虽然取得了突破，但仍然面临着一些挑战。例如，训练 DiT 模型需要消耗大量的计算资源和时间成本，这对于个人开发者和小型企业来说是一个很大的负担。此外，DiT 模型的内部机制比较复杂，目前对其工作原理的解释还不够充分，导致其可解释性不足。同时，DiT 技术的快速发展也引发了人们对于伦理和社会责任的担忧，例如，利用 DiT 技术生成虚假信息、侵犯个人隐私等问题也需要引起我们的重视。

未来的研究方向将集中在模型优化、效率提升、可解释性增强等方面。相信随着技术的不断进步，DiT 技术将在图像生成领域发挥越来越重要的作用，推动人工智能技术的发展和应用，为人类创造更多价值。
