

汇报主题：视觉生成的范式转移——从序列预测到尺度预测 (VAR)

1. Slide 1: 封面

【PPT内容】

- **标题：**Visual Autoregressive Modeling (VAR): Scalable Image Generation via Next-Scale Prediction
- **副标题：**视觉生成的“GPT时刻”与Scaling Laws的验证
- **来源：**NeurIPS 2024 (Best Paper Award)
- **汇报人：**[你的名字]

【演讲稿】

大家好，今天我要汇报的是一篇在 NeurIPS 2024 上获得最佳论文奖的重量级工作——VAR (Visual Autoregressive Modeling)。

在过去几年，我们见证了以 Stable Diffusion 为代表的扩散模型统治了AI绘画。但是，今天这篇论文提出了一个挑战性的问题：像 GPT 那样简单的“预测下一个词”的模式，能不能在画图上也打败扩散模型？

答案是肯定的。VAR 提出了一种全新的“下一尺度预测”范式，不仅速度比扩散模型快得多，更重要的是，它首次在视觉领域完美验证了类似 LLM 的 Scaling Laws (扩展定律)，这意味着视觉生成可能迎来了它的“GPT-4 时刻”。

2. Slide 2: 现有技术的痛点 (Motivation)

【PPT内容】

- **当前主流：扩散模型 (Diffusion Models)**
 - 优点：画质好、覆盖率高。
 - 缺点：推理慢（几十步迭代）、计算成本高、与LLM架构不统一。
- **传统自回归 (AutoRegressive, AR)**
 - 代表：VQGAN, DALL-E 1。
 - 逻辑：Next-Token Prediction (光栅扫描 Raster Scan)。
 - 致命伤：
 - 违反直觉：图像是二维的，不是一维序列。

2. **计算爆炸**: 生成一张图需要预测几千次, $O(n^6)$ 复杂度 1。

【演讲稿】

我们先看看背景。现在最火的文生图模型, 比如 Sora 或 Stable Diffusion, 大多是扩散模型。它们效果很好, 但有一个通病: 慢。生成一张图需要反复去噪几十次。

另一派是像 GPT 一样的自回归模型 (AR)。早期的 AR 模型 (如 VQGAN) 试图把图像切成一串像素点, 从左上角开始, 一行一行地“写”出图片。

VAR 的作者指出, 这种做法是反直觉的。人类看图或画图, 从来不是像打印机一样一行行扫描的, 而是先看整体轮廓, 再看细节。强行把二维图像拉成一维序列, 不仅破坏了空间结构, 还导致计算量随着分辨率爆炸式增长, 这直接限制了 AR 模型在高清图像上的表现。

3. Slide 3: 核心创新——下一尺度预测 (Next-Scale Prediction)

【PPT内容】

- **范式重构**:
 - **Old (AR)**: Next-Token Prediction (预测下一个像素)。
 - **New (VAR)**: Next-Scale Prediction (预测下一层清晰度)。
- **人类视觉逻辑**: Coarse-to-Fine (由粗到细)。
- **数学定义**:

$$p(Im) = \prod p(Scale_k | Scale_{<k})$$

- **关键优势**:
 - 层内并行生成 (Parallel generation within scale)。
 - 保留空间局部性 (Spatial Locality)。

【演讲稿】

VAR 的核心贡献, 就是把“预测下一个像素”改成了**“预测下一层清晰度”。

请看这张对比图。VAR 将图像编码成不同分辨率的特征图金字塔 (比如 16×16 到 32×32 再到 256×256)。

模型生成图片的过程变成了: 先生成一个模糊的 16×16 缩略图, 以此为基础, 预测 32×32 的细节残差, 以此类推, 直到生成高清原图。

这种 Coarse-to-Fine (由粗到细) 的策略有两个巨大的好处:

第一, 它符合人类视觉原理;

第二, 在生成每一层 (Scale) 时, 所有的 Token 是并行计算**的。这意味着它不需要像老方法那样预测 65,536 次, 而是只需要预测不到 10 次 (层数) 就能完成生成 1。

4. Slide 4: 复杂度分析 (Efficiency)

【PPT内容】

- 计算复杂度对比:

- 传统 AR: $O(n^6)$ —— 分辨率翻倍, 计算量暴增64倍。
- VAR: $O(n^4)$ —— 分辨率翻倍, 计算量增加16倍 1。

- 推理速度:

- 比传统 VQGAN 快 ~20倍。
- 比 DiT (Diffusion Transformer) 快 ~45倍 (在同等采样步数下) 1。

【演讲稿】

为了证明这种方法的优越性, 作者做了数学推导。

传统的自回归模型, 随着图像分辨率 n 的提升, 计算复杂度是可怕的 $O(n^6)$ 。这也是为什么以前的 AR 模型很难做高清图。

而 VAR 将其降维到了 $O(n^4)$ 。这在工程上是巨大的胜利。

实测数据显示, 在生成 256×256 图像时, VAR 的速度是传统 VQGAN 的 20 倍。哪怕对比现在最先进的扩散模型 DiT, VAR 在速度上也具有碾压级的优势 1。

5. Slide 5: 视觉领域的 Scaling Laws (重点)

【PPT内容】

- 什么是 Scaling Laws?

- 模型越大、算力越多 → 效果越好 (且可预测) 。
- LLM 成功的基石 (Chinchilla Laws)。

- VAR 的发现:

- 完美拟合: Test Loss 与参数量呈幂律关系。
- 相关系数: Pearson ≈ -0.998 (极度平滑的线性关系) 1。

- 意义: 证明了视觉生成可以像 LLM 一样通过堆算力持续变强。

【演讲稿】

这一页是本篇论文最震撼的部分, 也是它获得 Best Paper 的主要原因。

在 GPT 出现之前, 我们知道模型越大越好, 但不知道具体好多少。Scaling Laws 告诉我们, 性能提升是可以被数学预测的。

VAR 第一次在视觉生成领域画出了这条完美的曲线。大家看这张图 (指向 Scaling Law 曲线), 随着模型参数从 18M 增加到 2B, 测试集的 Loss 呈现出近乎完美的线性下降, 相关系数高达 -0.998 1。

这意味着什么? 这意味着 VAR 还没有遇到瓶颈。如果我们像训练 GPT-4 一样, 给它百亿参数、万亿数据, 它的生成质量在理论上会持续提升。这是扩散模型目前较难呈现出的特性 2。

6. Slide 6: 实验结果对比 (Results)

【PPT内容】

- **ImageNet 256x256 Benchmark:**
 - **FID (图像质量):** 1.73 (VAR) vs 2.27 (DiT-XL/2)。
 - **IS (多样性):** 350.2 (VAR) vs 278.2 (DiT)。
- **结论:**
 - 历史上首次 AR 模型在图像质量上**超越扩散模型**。
 - 同时保持了极高的推理速度。

【演讲稿】

口说无凭，跑分见真章。在标准的 ImageNet 评测中，20亿参数的 VAR 模型取得了 FID 1.73 的成绩。

FID 越低代表图像越真实。此前，这个榜单一直被扩散模型 (DiT) 霸榜。VAR 不仅打破了 AR 模型“生成质量差”的刻板印象，更是在分数上直接超越了 Sora 的基础架构 DiT。

这证明了：只要找对了序列建模的方式（即下一尺度预测），自回归模型完全可以生成比扩散模型更逼真的图像。

7. Slide 7: 零样本泛化 (Zero-Shot Capability)

【PPT内容】

- **涌现能力 (Emergent Abilities):**
 - **In-painting (补全):** 根据周围环境补全缺失部分。
 - **Out-painting (外推):** 向外扩展图像。
 - **Editing (编辑):** 修改特定内容。
- **关键点：**这些任务**没有**经过专门训练，是模型“学”会的 1。

【演讲稿】

除了画图，VAR 还展现了类似 GPT 的“通用智能”。

在没有针对性微调的情况下，VAR 可以直接做图像补全和扩充。比如把这张图中间挖空，VAR 能根据上下文理解，填补出符合逻辑的物体。

这说明 VAR 不是在死记硬背像素，而是真正理解了图像的全局结构和语义信息。这种 Zero-shot 能力是通向通用视觉模型 (AGI) 的重要标志。

8. Slide 8: 最新进展与未来 (Future Works)

【PPT内容】

- **VAR 的进化版:**
 - **Infinity (2024.12):** 引入 **Bitwise Modeling** (按位预测) , 突破分辨率限制 (4K+)，速度是 SD3 的 2.6倍。
 - **InfinityStar (2025.11):** **Text-to-Video**，全能多模态模型，生成 720p 视频比扩散模型快 10 倍。
- **总结论:**
 - VAR 开启了视觉自回归的新赛道。
 - 多模态统一 (Unification) 正在加速到来。

【演讲稿】

最后, VAR 并不是终点, 而是一个新时代的起点。

就在 VAR 发布后的几个月内, 其团队已经推出了更强的后续版本。

一个是 Infinity, 它通过“按位预测”解决了超高分辨率生成的问题, 现在是世界上最快的文生图模型之一。

另一个是刚刚发布的 InfinityStar, 它将 VAR 的逻辑扩展到了视频生成, 实现了图像和视频生成的模型统一。

这些进展都在告诉我们: 视觉生成的未来, 很可能属于这种可扩展的、统一的自回归架构。我的汇报到此结束, 谢谢大家!

9. Slide 9: Q&A (备选问题库)

(为应对同学或老师提问准备)

- **Q: 既然 VAR 这么好, 为什么现在大家还在用 Stable Diffusion?**
 - A: 生态惯性。SD 有庞大的社区和插件 (ControlNet, LoRA) 。VAR 是新技术, 生态还在建设中, 但技术潜力 (Scaling Law) 更高。
- **Q: VAR 和 VQGAN 到底区别在哪?**
 - A: 核心是“顺序”。VQGAN 是从左上到右下一个一个猜; VAR 是从模糊到清晰一层一层猜。后者利用了并行计算, 所以快得多 2。