释放生成对抗网络在非自回归文本生成中的潜力

Da Ren
计算学系 香港理工大学 中国香港
csdren@comp.polyu.edu.hk

Yi Cai大数据与智能机器人重点实验室（华南理工大学） 中国教育部 软件工程学院 华南理工大学 中国广州 ycai@scut.edu.cn

Qing Li计算学系 香港理工大学 中国香港 csqli@comp.polyu.edu.hk

arXiv:2305.03977v3 [cs.CL] 2 Oct 2024

摘要

生成对抗网络（GAN）已被用于文本生成，以解决曝光偏差问题。尽管生成对抗网络取得了显著进展，但它们采用自回归结构，因此在训练和推理阶段都存在较高的延迟。虽然生成对抗网络可以通过采用非自回归（NAR）结构实现高效生成，但在非自回归模型中的探索极为有限。在这项工作中，我们率先研究了如何基于非自回归结构构建语言生成对抗网络。我们发现了制约基于生成对抗网络的非自回归模型性能的两个问题。首先，现有的潜变量引入方法会产生高度相似的表征，无法描述句子中不同词语的多样性。为解决这一问题，我们提出了位置感知自调制，从而提供更多样、更有效的表征。其次，在生成对抗网络训练不稳定的情况下，Transformer 中的注意力机制无法准确地建立词语依赖关系，因此我们采用依赖前馈网络来增强模型的依赖关系建模能力。借助这两项技术，我们提出了一种基于生成对抗网络的非自回归模型——对抗式非自回归 Transformer（ANT）。实验结果表明，ANT 仅通过单次前向传播即可达到与主流模型相当的性能，并且在潜在插值和半监督学习等各种应用中具有巨大潜力。

1 引言

生成对抗网络（GAN）（Goodfellow 等人，2014年）是一类强大的生成模型。不同于需要多步推理的扩散模型（Ho等人，2020年；Saharia等人，2022）和自回归（AR）模型（Zhang等人，2021年；Yu 等人，2022），生成对抗网络能够通过单次前向传播生成高质量样本，因此延迟要低得多（Sauer 等人，2023；Kang 等人，2023）。生成对抗网络最初被应用于文本生成，以解决众所周知的曝光偏差问题（Bengio 等人，2015年）。该问题源于基于最大似然估计（MLE）的自回归模型在训练过程与推理过程之间存在差异。这些语言生成对抗网络（Yu 等人，2017；de Masson d' Autume 等人，2019年；Ren & Li，2023）保留了自回归结构，并通过在训练和推理阶段均使用先前生成的词作为输入来解决曝光偏差问题。在保证训练和测试过程一致的同时，它们不支持并行计算，因此延迟较高。图像生成对抗网络的高效特性在现有语言生成对抗网络中完全丧失了。从理论上讲，当且仅当学习到的分布与真实分布完全相同时，生成对抗网络才能达到全局最优性（Goodfellow 等人，2014年）。更

重要的是，它们的收敛并不依赖于特定结构。生成对抗网络理论上可以通过单次前向传播获得高质量样本。即便如此，它们在非自回归（NAR）文本生成方面的探索仍极其有限。

本文发掘了生成对抗网络在构建非自回归文本生成模型方面的潜力。不同于现有语言生成对抗网络所采用的自回归结构（Yu 等人，2017；Lin 等人，2017；Che 等人，2017；de Masson d'Autume 等人，2019；Ren & Li，2023），我们的模型采用支持高效并行计算的非自回归结构。在享受生成效率提升的同时，我们观察到现有自回归语言生成对抗网络与非自回归语言生成对抗网络之间存在明显的性能差距。主要障碍来自两个问题。首先，生成对抗网络依赖潜变量（其始终来自预定义分布）来支持采样，而现有引入潜变量的方法会产生高度相似的表征。这些表征无法描述词语之间的多样性，从而导致句子生成不准确。其次，Transformer（Vaswani 等人，2017）广泛应用于非自回归模型（Gu 等人，2018；Ghazvininejad 等人，2019）中，仅通过注意力机制建立词语依赖关系。然而，在生成对抗网络脆弱的训练过程中，动态权重分配过程会变得不稳定，导致词语依赖关系丢失，并最终产生不符合语法的输出。

针对第一个问题，我们提出了**位置感知自调制**，它可以为模型提供多样化的隐藏表征，从而生成句子中的不同词语。针对第二个问题，我们将 Transformer 中原有的前馈网络（FFN）模块替换为我们提出的**依赖前馈网络（依赖前馈网络）**。与注意力机制不同，注意力机制的依赖关系容易因权重分配不佳而丢失，而依赖前馈网络为依赖关系建模提供了更稳定的方法。借助这两项技术，我们提出了**对抗式非自回归Transformer（ANT）**。本工作的贡献总结如下：

- 我们率先开展了在非自回归结构中构建语言生成对抗网络的工作。为了支持生成句子中的不同词语，我们提出了位置感知自调制，它能够获得多样化的表征，以描述不同词语的多样性并提升生成质量。此外，我们提出了依赖前馈网络，以支持更稳定的依赖关系建模。与注意力机制不同，注意力机制容易受到生成对抗网络训练过程不稳定的影响，而依赖前馈网络能够帮助模型建立更准确的依赖关系，从而生成语法更通顺的结果。

- 利用这两项特性，我们提出了一种基于生成对抗网络的非自回归文本生成模型 ANT。现有的语言生成对抗网络采用自回归结构，因此依赖此前生成的词语，导致延迟较高。而 ANT 可以并行生成所有词语，实现高效生成。它继承了生成对抗网络的优势，能够通过单次前向传播生成高质量样本。
- 实验结果表明，在无条件生成和条件生成任务中，ANT 的性能与现有模型相当，但延迟显著更低。此外，我们还探索了 ANT 在潜在插值和半监督学习等应用中的潜力。据我们所知，这是首项展示生成对抗网络在构建非自回归文本生成模型方面有效性的研究。

2 背景

生成对抗网络（Goodfellow 等人，2014年）最初被引入文本生成领域，以解决臭名昭著的曝光偏差问题（Bengio 等人，2015年）。具体而言，现有文本生成模型采用自回归结构作为骨干，并以最大似然估计（MLE）作为训练目标。这些方法在训练时使用真实数据作为输入，但在推理时读取此前生成的词语。当模型在生成过程中出错时，这些错误会作为输入反馈给模型，使模型进入训练期间从未遇到过的状态空间（Bengio 等人，2015年）。因此，这些错误会迅速累积放大，导致生成样本的质量急剧下降。

生成对抗网络的训练不需要以真实数据作为输入，因此在训练和推理阶段都可以使用生成的词元。它通过在训练和测试过程中采用一致的生成方式，解决了曝光偏差问题。在文本生成中，生成器通常会对……进行建模。

输出词语概率，并从这些概率中采样特定词语。然而，这一采样操作不可微，会阻止梯度传递到生成器。

早期研究通过使用 REINFORCE（Yu 等人，2017；Lin 等人，2017；Che 等人，2017；Guo 等人，2018；de Masson d'Autume 等人，2019）或连续松弛（Nie 等人，2019）来解决这一问题。然而，REINFORCE（Williams，1992）的方差较高，而 Gumbel-softmax（Jang 等人，2017）等连续松弛方法则是有偏估计器。基于这两种方法的模型依赖预训练技术来获得可接受的性能（Ren & Li，2023）。

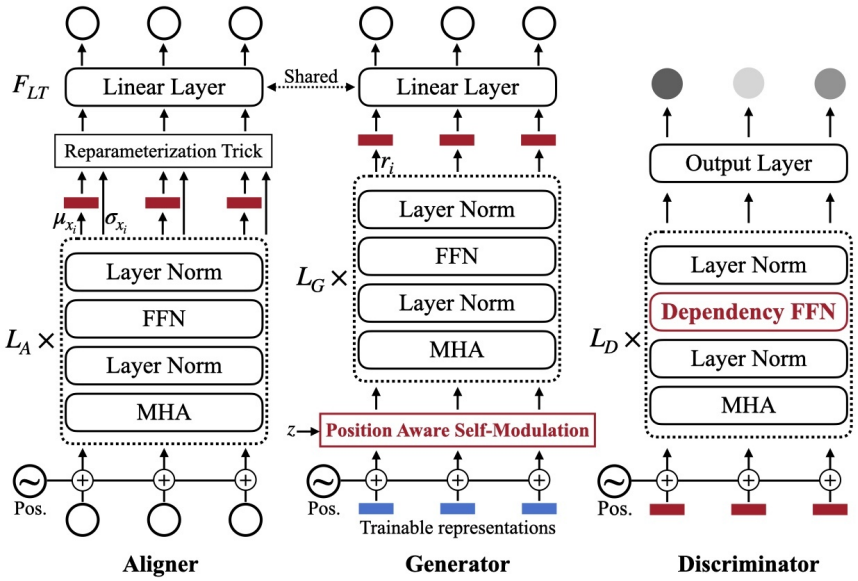

图 1：对抗式非自回归 Transformer（ANT）的结构

另一种方法是将词转换为表征，并训练生成器以获得这些表征。这种方法避免了训练过程中的不可微采样操作，因此梯度可以直接传递给生成器。与直接对词概率进行建模的 REINFORCE 和连续松弛方法不同，这些方法对词表征进行建模，因此被称为表示建模方法（Ren & Li，2023）。这种方法无需使用任何预训练技术，也能获得令人满意的性能。因此，我们基于这种方法构建模型。

NAGAN（Huang 等人，2021年）也采用生成对抗网络来构建非自回归模型，是与我们最相近的工作，但其模型的性能受到有偏直通估计器（Bengio 等人，2013）的显著限制，而我们的模型不受此问题影响。此外，我们提出的组件：位置感知自调制和依赖前馈网络能够进一步提升模型性能，而此前的工作并未探索这些组件。

此外，现有的基于最大似然估计的非自回归模型（Gu 等人，2018；Ghazvininejad 等人，2019）存在多模态问题 Gu 等人（2018），该问题倾向于将不同候选中的词语混合在一起，从而产生不符合语法的结果。Huang 等人（2022a）揭示了，这些模型学习到的分布与真实分布之间的 KL 散度仍然存在非负下界。从理论上讲，除非句子中的词语彼此独立（这与真实情况不符），否则学习到的分布不可能与真实分布完全相同。因此，现有的非自回归模型主要针对若干特定任务进行开发，而对更通用任务（如无条件生成）的探索极为有限。这进一步凸显了研究更有前景的方法以构建非自回归模型的迫切必要性，从而使其能够应用于多种领域。

3 模型

3.1 模型结构

本文提出了一种对抗式非自回归 Transformer（ANT），以完全非自回归的方式生成文本。ANT 基于表征建模框架（Ren & Li，2023）。如图 1 所示，ANT 包含三个部分：对齐器、判别器和生成器。对齐器将词语映射为表征，生成器则尝试恢复这些表征。判别器需要判断输入表征来自对齐器还是生成器。我们采用 Transformer（Vaswani 等人，2017）作为这三个部分的骨干网络，以支持高度并行的计算。输入首先与位置编码相加，然后送入编码器层。每个编码器层都包含一个多头注意力（MHA）模块和一个前馈网络（FFN）模块。每个模块之后都添加了层归一化。

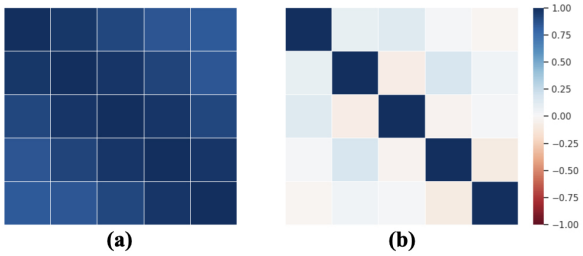

图2：自调制 (a) 和位置感知自调制 (b) 输出的余弦相似度。

对齐器经过训练，可根据掩码输入重构单词，这与 BERT 的训练过程相同（Devlin 等人，2019年）。沿用先前采用表示建模方法训练语言生成对抗网络的工作（Ren & Li，2023），我们使用变分自编码器（VAE）的损失函数训练对齐器（Kingma & Welling，2014年）：

$$L_A = -\mathbb{E}_{z_i' \sim q(z_i'|x_i)}(logp(x_i|z_i')) + KL(q(z_i'|x_i)||p(z_i')) \quad (1)$$

其中，x_i 是句子中的第 i 个单词，z_i' 通过重参数化技巧得到：$z_i' = \mu_{x_i} + \sigma_{x_i} \cdot \mathcal{N}(0, 1)$，$z_i'$则通过线性变换层 F_{LT}转换回单词。与将单词映射到表征空间中特定点的交叉熵不同，这种方法为每个单词描述一个区域，因此，即使表征稍微偏离其中心点 μ_{x_i}，仍然可以转换为正确的单词。

非自回归生成器无法输入先前生成的词，因此采用可训练的表征作为输入。随后，生成器在不同位置输出表征 r_i，并使用对齐器中的同一线性变换层 F_{LT} 将这些表征转换回词语。判别器将对齐器和生成器的输出表征（μ_{x_i} 和 r_i）作为输入。与图像生成对抗网络的判别器针对一幅图像输出单个标量不同，我们的判别器会针对每个表征给出输出。在训练期间，首先训练对齐器；在训练判别器和生成器时，其参数保持固定。训练过程中，生成器给出的表征无需转换为词语，因此判别器的梯度可以直接反向传播到生成器。

判别器和生成器都采用因果掩码，以打破输入中可能存在的对称性。我们使用 Wasserstein 距离（Arjovsky 等，2017）作为训练目标，并采用 Lipschitz 惩罚项（Petzka 等，2018）对判别器进行正则化。不过，我们的基础模型与现有自回归模型之间仍存在差距，因此进一步提出位置感知自调制和依赖前馈网络（依赖前馈网络）来提升模型性能。

3.2 位置感知自调制

有效的采样方法在生成对抗网络的成功中发挥着关键作用。基于 Transformer 的图像生成对抗网络（Lee 等，2021）通过采用自调制（Chen 等，

2019年）以融入潜变量。自调制为不同位置的归一化结果分配相同的平移和缩放因子，因此即使加入位置编码，不同位置的表征也会高度相似（如图2 (a)所示）。然而，生成器的输出（即不同位置的词表示）具有很高的多样性。相似的输入表征无法描述不同词语之间的多样性，从而导致句子生成不准确。

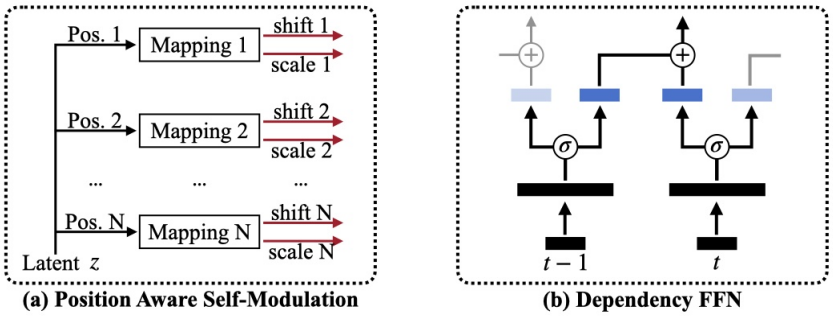

图3：ANT 中提出的方法。

为解决这一问题，我们提出**位置感知自调制**。如图3（a）所示，该方法针对不同位置的计算采用不同的映射层，从而获得多样化的结果。实际应用中，我们采用并行实现来提高计算效率，具体如下：

$$\begin{pmatrix} \mathbf{h}'_1 \\ \mathbf{h}'_2 \\ \vdots \\ \mathbf{h}'_N \end{pmatrix} = MLP(z)$$
$$\mathbf{h}_i = \gamma(\mathbf{h}'_i) \circ LN(\mathbf{x}_i) + \beta(\mathbf{h}'_i) \tag{2}$$

其中， z是潜变量，$\mathbf{h}'_i$是第 i个位置的隐藏表征， $MLP(\cdot)$是非线性变换，其激活函数为 GELU（Hendrycks & Gimpel，2016）， $LN(\cdot)$是层归一化， N 是句子长度， $\gamma(\cdot)$和 $\beta(\cdot)$是线性变换。在位置感知自调制中，不同位置的表征基于独有参数计算，彼此差异明显（如图2（b）所示），从而提供更有效的信号来生成目标句子。

3.3 依赖前馈网络

Transformer（Vaswani 等人，2017）通过在注意力机制中动态分配权重来建立词语依赖关系。然而，在生成对抗网络的训练过程中，这一过程并不稳定，导致模型丢失词语依赖关系，最终生成不符合语法的句子。为解决这一问题，我们提出依赖前馈网络（依赖前馈网络），以增强前馈网络模块的依赖关系建模能力。依赖前馈网络的结构如图3（b）所示，计算方式如下：

$$\begin{aligned} \mathbf{s}_t &= \sigma(\mathbf{x}_t W_s + b_s) \\ \mathbf{o}_t &= \mathbf{s}_{t-1} W_a + \mathbf{s}_t W_b + b_o \end{aligned} \tag{3}$$

其中， $\sigma(\cdot)$ 是一种激活函数，本研究中采用 GELU。使用因果掩码后，$\mathbf{s}_{t-1}$ 和$\mathbf{s}_t$分别包含第一个（$t-1$）和 t 个词的信息。在生成对抗网络脆弱的训练过程中，将这两个变量相加，有助于模型显式地建立第 t个词与前面（$t-1$）个词之间稳定的依赖关系。

3.4 条件生成的扩展

除了无条件生成外，条件生成也广泛应用于各种任务。因此，我们也将 ANT 扩展到条件生成。给定条件表示 c，

Ta表 1：COCO 数据集和 EMNLP 数据集上的 FED 和 I. BLEU（DI：解码迭代次数）。

模型	DI	COCO 数据集		EMNLP 数据集	
		FED ↓	I. BLEU ↑	FED ↓	I.BLEU ↑
训练数据	-	0.007	35.36	0.010	20.62
Transformer	O(N)	**0.008**	**34.28**	**0.014**	**19.50**
SeqGAN	O(N)	0.134	22.34	0.210	9.90
RankGAN	O(N)	0.203	22.10	0.290	10.37
MaliGAN	O(N)	0.074	25.95	0.079	13.11
LeakGAN	O(N)	0.132	29.43	0.125	11.59
RelGAN	O(N)	0.062	29.53	0.136	14.74
ScratchGAN	O(N)	0.014	30.76	0.018	17.19
InitialGAN	O(N)	0.013	33.06	0.025	17.74
CMLM	O(k)	0.016	27.65	0.062	**16.67**
NAT	O(1)	0.024	26.41	0.111	11.38
NAGAN	O(1)	0.084	24.98	0.748	2.01
ANT	O(1)	**0.013**	**31.12**	**0.026**	15.51

生成器可以通过偏移原始潜变量 z来考虑该条件。我们发现，使用可训练因子为 z 和 c 分配权重可以略微提升性能：$\hat{z} = \alpha_1 \circ z + \alpha_2 \circ c$，其中 α_1和 α_2 是两个可训练变量。对于判别器，我们将词表示x_t^d与条件表示 c 相加作为输入：$\hat{x}_t^d = x_t^d + c$。随后，将$\hat{x}_t^d$输入判别器的其余模块。

4 实验

4.1 实验设置

该实验涵盖无条件生成和条件生成，以全面评估模型性能。对于无条件生成，任务是生成分布尽可能接近目标集的句子。我们遵循先前工作的设置（de Masson d'Autume 等人，2019年；Ren & Li，2023），并使用来自两个数据集的句子：COCO 图像描述数据集（Lin 等人，2014年）[1] 和 EMNLP 2017 新闻数据集[2]。COCO 数据集和 EMNLP 数据集的训练集规模分别设为 50,000 和 200,000。COCO 数据集可用于评估短句子生成，而 EMNLP 数据集则侧重于长句子生成。对于条件生成，我们从 Yelp 数据集[3] 中随机选取 100,000 个句子作为训练数据，并使用情感标签（正面或负面）作为条件。

4.2 评估指标

评估在嵌入级别和词元级别进行。在嵌入级别，我们使用 Universal Sentence Encoder[4] (Cer 等人，2018)将句子转换为嵌入。然后，我们分别计算**Fr´echet Embedding Distance (FED)** (de Masson d'Autume 等人，2019)和**最低覆盖率 (LCR)** (Ren & Li，2023年)，以评估两个分布的整体相似度和细粒度相似度。

在词元级别，我们使用**逆向 BLEU（I. BLEU）**，同时从质量和多样性方面评估模型性能。此外，我们还通过调整模型温度，绘制**BLEU** （Papineni 等人，2002）和**Self-BLEU**（Zhu 等人，2018）的曲线（Caccia 等人，2020）。在这种情况下

[1]https://cocodataset.org[2]http://www.statmt.org/wmt17/[3]
https://www.yelp.com/dataset[4]
https://tfhub.dev/google/universal-sentence-encoder/4

在条件生成中，**准确率（准确率）**也用于评估模型生成的句子是否与输入标签一致。

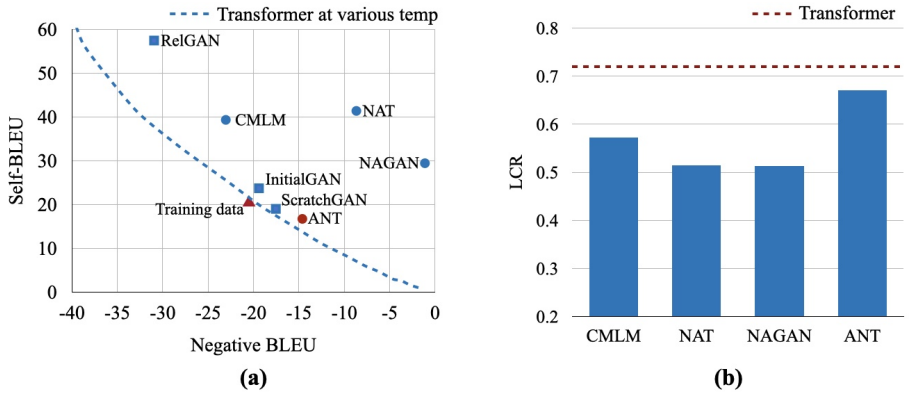

图 4：其他实验结果。(a) 模型在不同温度下的性能。(b) 最低覆盖率。

4.3 对比模型

一个重要的对比模型是 Transformer，它采用自回归结构，并通过最大似然估计进行训练。它是各种文本生成任务中的主流模型。此外，我们还对比了多个自回归语言生成对抗网络：基于*REINFORCE*的 SeqGAN（Yu 等人，2017）、RankGAN（Lin 等人，2017）、MaliGAN（Che 等人，2017）、LeakGAN（Guo 等人，2018）和 ScratchGAN（de Masson d'Autume 等人，2019）；使用 *Gumbel-softmax* 获取梯度的 RelGAN（Nie 等人，2019）；以及不采用上述两种方法、而采用表征建模的 InitialGAN（Ren & Li，2023）。上述所有模型都是自回归模型，其解码迭代次数（DI）为 $O(N)$（N 为序列长度）。

对于非自回归模型，我们将其与另一种基于 GAN 的模型 NAGAN（Huang 等人，2021年）进行比较。此外，我们的实验还比较了两种经典的基于最大似然估计的非自回归模型：非自回归 Transformer（NAT）（Gu 等人，2018）和条件掩码语言模型（CMLM）（Ghazvininejad 等人，2019年）。关于实验细节的更多说明见附录 A。我们将在未来公开代码。

4.4 实验结果

4.4.1 无条件生成

无条件生成的实验结果如表 1 所示。在 COCO 数据集上，Transformer 在自回归模型中表现最佳，而 ANT 在非自回归模型中表现最佳。具体而言，ANT 的 FED 得分为 0.013。这一结果优于多种自回归语言生成对抗网络，并接近在 COCO 数据集上表现最好的语言生成对抗网络 InitalGAN。在逆向 BLEU（I. BLEU）指标上也能观察到类似结果。ANT 的 I. BLEU 得分为 31.12，远优于其他非自回归模型。基于最大似然估计的非自回归模型（NAT 和 CMLM）的表现远逊于自回归模型。现有的基于最大似然估计的非自回归模型利用特定任务的特点（例如机器翻译中输入与输出之间的强对应关系）来缓解多模态问题 Gu 等人（2018）。一旦将它们应用于更基础的任务（例如无条件生成），这一固有问题就会更加严重，导致模型性能明显下降。另一种基于生成对抗网络的非自回归模型 NAGAN 则逊于所有其他模型。这表明存在有偏直通估计器的局限性。

对于 EMNLP 数据集，Transformer 仍然是表现最佳的模型。ANT 在 FED 上优于其他非自回归模型，而 CMLM 在逆向 BLEU 上略胜于 ANT。迭代解码机制有助于 CMLM 更好地处理复杂数据集，但解码延迟较高。为了进一步讨论它们在词元级别上的表现，我们遵循 Caccia 等人（2020）的建议，并

表 2：Yelp 数据集上的 FED、I. BLEU 和准确率

模型	DI	FED	I. BLEU	Acc.
训练数据	-	0.008	24.18	92.47%
Transformer	O(N)	0.011	23.04	91.73%
CMLM	O(k)	**0.015**	18.35	87.85%
NAT	O(1)	0.032	11.81	83.54%
ANT	O(1)	0.018	**19.08**	**88.35%**

通过调整 Transformer 的温度，绘制自 BLEU 和负 BLEU 曲线。结果如图 4（a）所示。ANT 是唯一能够达到与自回归模型相当表现的非自回归模型，而其他非自回归模型（包括 CMLM）的表现则明显落后。具体而言，NAGAN 的 BLEU 极低，这表明 NAGAN 无法生成流畅的句子，也揭示了 NAGAN 难以在复杂数据集上收敛。此外，我们在图 4（b）中比较了 Transformer 与其他非自回归模型的最低覆盖率（LCR）。ANT 以更少的解码迭代次数胜过 CMLM，也是唯一一个表现接近 Transformer 的非自回归模型。

4.4.2 条件生成

条件生成的实验结果如表 2 所示。在非自回归模型中，ANT 在 FED 上的表现与 CMLM 相当，并且在逆向 BLEU 和准确率方面更高，同时解码延迟更低。NAT 同样能在一个解码步骤中生成样本，但表现逊于其他模型。在准确率方面，ANT 达到 88.35%，是所有非自回归模型中最高的。ANT 能够生成与给定标签一致的句子。

无条件生成和条件生成中的实验结果均证明了 ANT 的有效性。它的表现优于多种自回归语言生成对抗网络。即使与现有表现最佳的语言生成对抗网络相比，ANT 仍能在显著降低延迟的情况下获得相当的性能。因此，没有必要采用自回归结构来构建语言生成对抗网络。此外，ANT 不受基于最大似然估计的非自回归模型在理论上的局限性，因此能够取得更好的性能，也代表了一种更有前景的非自回归模型构建方法。

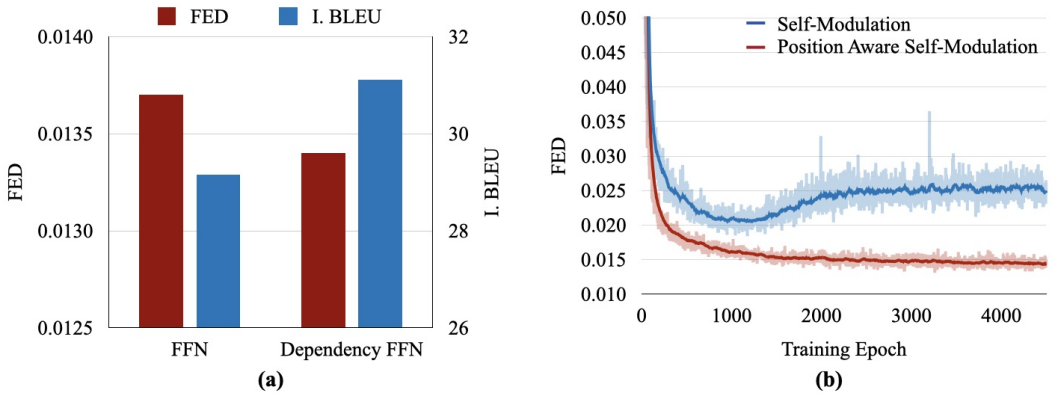

图5：对 (a) 依赖前馈网络和 (b) 位置感知自调制的消融研究。

4.4.3 消融研究

我们为 ANT 设计了两个特征：1）依赖前馈网络；2）位置感知自调制。我们通过实验展示这两个模块的有效性。对于依赖前馈网络，我们在图5 (a) 中比较了依赖前馈网络与原始前馈网络的性能。采用依赖前馈网络的 ANT 具有更低的 FED 和更高的逆向 BLEU，在词元级指标和嵌入级指标上都取得了更好的性能。这些结果表明，依赖前馈网络能够通过建模更准确的词语依赖关系来提升模型性能。

表 3：ANT 在半监督学习中的有效性（Num.：标注数据数量 ）.

方法	Num.	P	R	F1
SL	500	91.28%	89.06%	90.15%
SSL		90.77%	92.15%	**91.46%**
SL	1000	92.42%	91.33%	91.87%
SSL		94.87%	92.39%	**93.62%**

对于位置感知自调制，我们在图5 (b) 中将其训练曲线与采用 FED 的原始自调制进行了比较。采用位置感知自调制的 ANT 收敛速度更快，最终取得了更好的性能。位置感知自调制能够通过提供更多样化且有效的表征来提升模型性能。

4.5 讨论

ANT 的一个优势是只需一个解码步骤，且加速比很高。我们在图 6（a）中比较了不同模型的加速比。ANT 的速度比 Transformer 快 14.75 倍。即使与 CMLM 相比，ANT 的解码延迟也低得多，同时取得了相当甚至更好的性能。此外，ANT 在各种应用中都具有巨大潜力。接下来，我们将探讨 ANT 在半监督学习和潜在空间插值等不同应用中的潜力。

生成模型可以融入半监督学习（SSL），以辅助其他模型的训练。由于训练过程中需要获取新数据，因此模型必须具备较高的效率。我们通过将 ANT 融入分类模型的训练，探究其在 SSL 中的应用。该分类模型经过训练，可以识别 Yelp 数据集句子的情感标签。我们准备了两个训练集：一个包含 500 条标注数据，另一个包含 1,000 条标注数据。结果如表 3 所示。对于使用 500 条标注数据训练的分类模型，其 F1 分数从 90.15% 提升至 91.46%。使用 SSL 后，性能提升了 +1.31%。对于使用 1,000 条标注数据训练的模型，其 F1 分数从 91.87% 提升至 93.62%，其中 +1.75% 的提升归功于 SSL。使用 SSL 训练的分类模型始终优于使用监督学习（SL）训练的模型。ANT 能够生成与原始数据具有相同分布的数据，从而帮助分类模型提升性能。

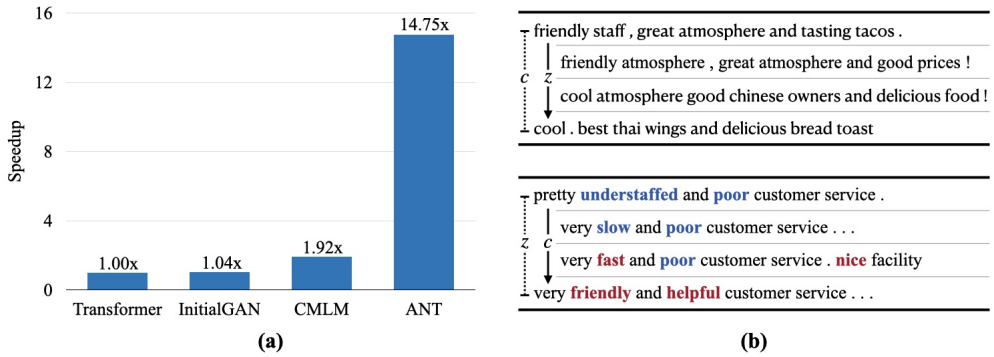

图 6：（a）不同模型的加速比。（b）潜在插值案例研究。

此外，ANT 也能像图像生成对抗网络一样进行潜在插值。ANT 中有两个潜变量：z，它从预定义分布中采样；以及 c，它是一种条件表示。我们固定其中一个，并逐步改变另一个。图 6 (b) 中的第一个示例展示了在固定 c 的情况下调整 z 得到的样本，其中 ANT 将一个句子转换为另一个句子，同时确保中间句子仍然通顺易懂。第二个示例展示了将 c 从负向表示变为正向表示时得到的样本。ANT 在保留句子主体结构的同时，逐步将负面词语转换为正面词语。这种

潜在插值很少在非自回归模型中得到探索，也可能为相关任务带来进一步的启发。

5 结论

在这项工作中，我们首先揭示了现有基于自回归结构的语言生成对抗网络的局限性。生成对抗网络的全局最优性并不依赖于特定结构，而非自回归结构能够以低得多的延迟取得相近的性能。随后，我们指出了构建基于生成对抗网络的非自回归模型时面临的两个问题：1）现有的潜变量整合方法得到的表征相似，无法描述不同词语的多样性。2）Transformer 中的注意力机制无法稳定地建立词语依赖关系。针对这两个问题，我们分别提出了位置感知自调制和依赖前馈网络。借助这两项技术，我们提出了对抗式非自回归Transformer（ANT），一种基于生成对抗网络的非自回归模型。实验结果表明，ANT 的性能可与其他自回归模型相媲美，同时解码延迟低得多。此外，我们还展示了 ANT 在平滑潜在插值和半监督学习等多种任务中的巨大潜力。

参考文献

Mart´ın Abadi、Ashish Agarwal、Paul Barham、Eugene Brevdo、Zhifeng Chen、Craig Citro、Greg S. Corrado、AndyDavis、JeffreyDean、Matthieu Devin、Sanjay Ghemawat、Ian Goodfellow、Andrew Harp、Geoffrey Irving、Michael Isard、Yangqing Jia、Rafal Jozefowicz、Lukasz Kaiser、Manjunath Kudlur、Josh Levenberg、Dandelion Man´e、Rajat Monga、Sherry Moore、Derek Murray、Chris Olah、Mike Schuster、Jonathon Shlens、Benoit Steiner、Ilya Sutskever、Kunal Talwar、Paul Tucker、Vincent Vanhoucke、Vijay Vasudevan、Fernanda Vi´egas、Oriol Vinyals、Pete Warden、Martin Wat- tenberg、Martin Wicke、Yuan Yu 和 Xiaoqiang Zheng。《TensorFlow：异构系统上的大规模机器学习》，2015年。网址：https://www.tensorflow.org/。软件可从 tensorflow.org 获取。

Mart´ın Arjovsky、Soumith Chintala 和 L´eon Bottou。Wasserstein 生成对抗网络。发表于 *ICML*，第 214–223 页，2017 年。网址：http://proceedings.mlr.press/v70/ arjovsky17a.html。

Samy Bengio、Oriol Vinyals、Navdeep Jaitly 和 Noam Shazeer。使用循环神经网络进行序列预测的计划采样。发表于*NeurIPS*，第 1171–1179 页，2015年。

Yoshua Bengio、Nicholas L´eonard 和 Aaron C. Courville。为条件计算估计或传播随机神经元的梯度。*CoRR*，abs/1308.3432，2013年。

Massimo Caccia、Lucas Caccia、William Fedus、Hugo Larochelle、Joelle Pineau 和 Laurent Char- lin。语言生成对抗网络表现不佳。发表于*ICLR*，2020年。网址：https://openreview.net/forum? id=BJgza6VtPB。

Daniel Cer、Yinfei Yang、Sheng-yi Kong、Nan Hua、Nicole Limtiaco、Rhomni St. John、Noah Con- stant、Mario Guajardo-Cespedes、Steve Yuan、Chris Tar、Brian Strope 和 Ray Kurzweil。英语通用句子编码器。在*EMNLP*，2018 年。doi：10.18653/v1/d18-2029。

Tong Che, Yanran Li, RuixiangZhang, R.DevonHjelm, Wenjie Li, YangqiuSong, andYoshua Bengio。最大似然增强型离散生成对抗网络。 *CoRR*，abs/1702.07983，2017。

Ting Chen、Mario Lucic、Neil Houlsby 和 Sylvain Gelly。《论生成对抗网络的自调制》。发表于 *ICLR*，2019年。网址：https://openreview.net/forum?id=Hkl5aoR5tm。

Cyprien de Masson d'Autume、Shakir Mohamed、Mihaela Rosca 和 Jack W. Rae。从头开始训练语言生成对抗网络。发表于*NeurIPS*，第 4302–4313 页，2019年。网址：https://proceedings.neurips.cc/paper/2019/hash/ a6ea8471c120fe8cc35a2954c9b9c595-Abstract.html。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。发表于*NAACL-HLT*，第 4171–4186 页，2019年。doi：10.18653/v1/n19-1423。

Marjan Ghazvininejad、Omer Levy、Yinhan Liu 和 Luke Zettlemoyer。Mask-predict：条件掩码语言模型的并行解码。发表于*EMNLP-IJCNLP*，第 6111–6120 页，2019年。doi：10.18653/v1/D19-1633。

Ian J. Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron C. Courville 和 Yoshua Bengio。生成对抗网络。载于 *NeurIPS*，第 2672–2680 页，2014年。

Jiatao Gu、James Bradbury、Caiming Xiong、Victor O. K. Li 和 Richard Socher。非自回归神经机器翻译。发表于 *ICLR*，2018 年。网址：https://openreview. net/forum?id=B1l8BtlCb。

Jiaxian Guo、Sidi Lu、Han Cai、Weinan Zhang、Yong Yu 和 Jun Wang。通过结合泄露信息的对抗训练实现长文本生成。载于*AAAI*，第 5141–5148 页，2018。

DanHendrycks 和 KevinGimpel。通过高斯误差线性单元衔接非线性函数与随机正则化器。*CoRR*，abs/1606.08415，2016年。

Martin Heusel、Hubert Ramsauer、Thomas Unterthiner、Bernhard Nessler 和 Sepp Hochreiter。采用双时间尺度更新规则训练的生成对抗网络收敛至局部纳什均衡。载于*NeurIPS*，第 6626–6637 页，2017。

Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。载于*NeurIPS*，2020年。

Fei Huang、Jian Guan、Pei Ke、Qihan Guo、Xiaoyan Zhu 和 Minlie Huang。用于语言生成的文本 { 生成对抗网络} 与非自回归生成器，2021。网址：https://openreview. net/forum?id=wOI9hqkvu_。

FeiHuang、Tianhua Tao、Hao Zhou、Lei Li 和 Minlie Huang。论非自回归变换器的学习。发表于*ICML*，第 9356–9376 页，2022年（a）。网址：https://proceedings.mlr. press/v162/huang22k.html。

Xiao Shi Huang、Felipe Perez 和 Maksims Volkovs。无需知识蒸馏即可改进非自回归翻译模型。发表于*ICLR*，2022年（b）。网址：https://openreview.net/forum? id=I2Hw58KHp8O。

Eric Jang、Shixiang Gu 和 Ben Poole。使用 Gumbel-softmax 进行类别重参数化。发表于*ICLR*，2017。

Minguk Kang、Jun-Yan Zhu、Richard Zhang、Jaesik Park、Eli Shechtman、Sylvain Paris 和 Taesung Park。扩展生成对抗网络以实现文本到图像合成。发表于*CVPR*，第 10124–10134 页。IEEE，2023。doi：10.1109/CVPR52729.2023.00976。

Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。发表于 *ICLR*，2015年。网址：http://arxiv.org/abs/1412.6980。

Diederik P. Kingma 和 Max Welling。自动编码变分贝叶斯。在 *ICLR*，2014年。网址 http://arxiv.org/abs/1312.6114。Kwonjoon Lee、Huiwen Chang、Lu Jiang、Han Zhang、Zhuowen Tu 和 Ce Liu。Vitgan：使用视觉 Transformer 训练生成对抗网络。*CoRR*，abs/2107.04589，2021年。Kevin Lin、Dianqi Li、Xiaodong He、Ming-Ting Sun 和 Zhengyou Zhang。面向语言生成的对抗性排序。在*NeurIPS*，第 3155–3165 页，2017。网址 https://proceedings.neurips.cc/paper/2017/hash/bf201d5407a6509fa536afc4b380577e-Abstract.html。Tsung-Yi Lin、Michael Maire、Serge J. Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C. Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。在 *ECCV*，第 740–755 页，2014年。doi：10.1007/978-3-319-10602-1\ 48。

Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。在*ICLR*，2019年。网址 https://openreview.net/forum?id=Bkg6RiCqY7。

Weili Nie、Nina Narodytska 和 Ankit Patel。《Relgan：用于文本生成的关系生成对抗网络》。发表于 *ICLR*，2019年。网址：https://openreview.net/forum?id=rJedV3R5tm。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。BLEU：一种机器翻译自动评估方法。发表于*ACL*，第 311–318 页，2002 年。doi：10.3115/1073083.1073135。

Henning Petzka、Asja Fischer 和 Denis Lukovnikov。关于 Wasserstein生成对抗网络的正则化。发表于*ICLR*，2018 年。网址：https://openreview.net/forum?id=B1hYRMbCW。

Da Ren 和 Qing Li。InitialGAN：一种采用完全随机初始化的语言生成对抗网络。*IEEE*神经网络与学习系统汇刊，第 1–14 页，2023 年。doi：10.1109/TNNLS。2023.3315778。

Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily L. Denton、Seyed Kamyar Seyed Ghasemipour、Raphael Gontijo Lopes、Burcu Karagol Ayan、Tim Salimans、Jonathan Ho、David J. Fleet 和 Mohammad Norouzi。具有深度语言理解能力的照片级真实感文本到图像扩散模型。在 Sanmi Koyejo、S.Mohamed、A. Agarwal、Danielle Belgrave、K. Cho 和 A. Oh（编）所编论文集，*NeurIPS*，2022 年。

Axel Sauer、Tero Karras、Samuli Laine、Andreas Geiger 和 Timo Aila。StyleGAN-T：释放生成对抗网络的力量，实现快速的大规模文本到图像合成。在 Andreas Krause、Emma Brunskill、Kyunghyun Cho、Barbara Engelhardt、Sivan Sabato 和 Jonathan Scarlett（编）所编论文集，*ICML*，第 202 卷，机器学习研究论文集，第 30105–30118 页。PMLR，2023 年。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。注意力就是一切。发表于 *NeurIPS*，第 5998–6008 页，2017 年。网址：https://proceedings.neurips.cc/paper/2017/hash/ 3f5ee243547dee91fbd053c1c4a845aa-Abstract.html。

Ronald J. Williams. 用于连接主义强化学习的简单统计梯度跟随算法。机器学习，8:229–256，1992。doi: 10.1007/BF00992696。

Jiahui Yu、Yuanzhong Xu、Jing Yu Koh、Thang Luong、Gunjan Baid、Zirui Wang、Vijay Vasudevan、Alexander Ku、Yinfei Yang、Burcu Karagol Ayan、Ben Hutchinson、Wei Han、Zarana Parekh、Xin Li、HanZhang、Jason Baldridge 和 Yonghui Wu。扩展自回归模型以生成内容丰富的文本到图像内容。机器学习研究汇刊，2022，2022。

Lantao Yu、Weinan Zhang、Jun Wang 和 Yong Yu。《SeqGAN：采用策略梯度的序列生成对抗网络》。载于*AAAI*，第 2852–2858 页，2017 年。

Han Zhang、Weichong Yin、Yewei Fang、Lanxin Li、Boqiang Duan、Zhihua Wu、Yu Sun、Hao Tian、Hua Wu 和 Haifeng Wang。《ERNIE-ViLG：用于双向视觉-语言生成的统一生成式预训练》。*CoRR*，abs/2112.15283，2021年。

Yaoming Zhu、Sidi Lu、Lei Zheng、Jiaxian Guo、Weinan Zhang、Jun Wang 和 Yong Yu。《Texygen：文本生成模型的基准测试平台》。载于*SIGIR*，第 1097–1100 页，2018 年。doi：10.1145/3209978.3210080。

A 实验细节

A.1 实现细节

对齐器、生成器和判别器的层数均设为 4。它们的输入维度为 256，前馈网络 / 依赖前馈网络的隐藏维度为 1,024。头数设为 8。我们使用 AdamW（Loshchilov 和 Hutter，2019年）作为对齐器的优化器，权重衰减设为 1e-5；其学习率为 0.0001。在对抗训练期间，使用 AdamW（Loshchilov 和 Hutter，2019年）作为判别器的优化器，其权重衰减设为 0.0001；

对于 COCO 和 Yelp 数据集，其学习率为 0.0002；对于 EMNLP 数据集，学习率为 0.00015。我们选择 Adam（Kingma 和 Ba，2015年）作为生成器的优化器，其学习率为 0.0001。判别器和生成器优化器中的 β_1 和 β_2 分别设为 0.5 和 0.9。最大训练轮数设为 4,500。我们基于 Tensorflow[5]（Abadi 等，2015年）实现模型，并在 NVIDIA GeForce RTX 3090 上训练模型。

A.2 NAT 和 CMLM 的实现

NAT（Gu 等人，2018）和 CMLM（Ghazvininejad 等人，2019）都是为机器翻译而设计的，因此其原始结构包含一个用于编码源语言输入的编码器。然而，在我们的实验中可能没有有意义的输入（例如，无条件生成），因此无法在实验中直接使用这种结构。因此，我们借鉴 VAE 的思路来获得隐层表征，使其能够迁移到我们的实验任务中。
更具体地说，训练期间采用基于 Transformer 的编码器将句子编码为隐层表征。随后，将这些表征输入解码器，以重构输入句子。推理期间，从标准正态分布中采样得到的表征将输入解码器，解码器则基于采样得到的表征生成句子。对于 NAT，表征作为输入送入解码器。对于 CMLM，表征与输入词元（被掩码或未被掩码的词语）的嵌入拼接在一起。与先前的工作（Ghazvininejad 等人，2019；Huang 等人，2022年（b））一样，CMLM 的迭代次数设为 10。

A.3 评估指标

Fr´echet 嵌入距离（FED）（de Masson d'Autume 等人，2019年）与 Fr´echet Inception Distance（FID）（Heusel 等人，2017）相同，区别仅在于编码模型。编码模型经过调整，以适用于文本生成。我们按照先前工作的设置（de Masson d'Autume 等人，2019年）采用了通用句子编码器。其计算方式如下：

$$FED = ||\mu_1 - \mu_2||_2^2 + Tr(c_1 + c_2 - 2(c_1 c_2)^{1/2}) \tag{4}$$

其中， μ_1 和 μ_2 是均值， c_1 和 c_2 是协方差。

最低覆盖率（LCR）（Ren & Li，2023）被提出作为补充指标，用于在两个模型的 FED 非常接近时进行评估，因为 LCR 对数据质量的变化更加敏感（Ren & Li，2023）。给定两组句子 $X_i^a \in \mathbb{X}_a$ 和 $X_i^b \in \mathbb{X}_b$，LCR 的计算方式如下：

$$\begin{aligned} S_{ij} &= Sim(\mathbf{Emb}(X_i^a), \mathbf{Emb}(X_j^b)) \\ R_a &= \frac{1}{|\mathbb{X}_a|} \sum_{i=1}^{|\mathbb{X}_a|} \delta(\sum_{j=1}^{|\mathbb{X}_b|} S_{ij} \geq \tau) \\ R_b &= \frac{1}{|\mathbb{X}_b|} \sum_{j=1}^{|\mathbb{X}_b|} \delta(\sum_{i=1}^{|\mathbb{X}_a|} S_{ij} \geq \tau) \\ LCR(\mathbb{X}_a, \mathbb{X}_b) &= min(R_a, R_b) \end{aligned} \tag{5}$$

其中， X_i^a 和 X_i^a 分别是句子集 $\mathbb{X}_a$ 和 $\mathbb{X}_b$中的第 i 个和第 j 个句子。$\mathbf{Emb}(\cdot)$是用于将句子转换为嵌入向量的模型（本研究中使用的是 Universal Sentence Encoder）， τ 是一个超参数， $Sim(\cdot)$表示余弦相似度， $\delta(\cdot)$是一个函数：当输入大于 0 时返回 1，否则返回 0。

LCR 的思路是判断一个集合中的某个特定模式是否被另一个集合中的句子覆盖。随后，它将最小覆盖率作为输出，因此 LCR 对文本生成模型中的两个常见问题很敏感：1）模型倾向于生成偏离真实分布的句子；2）生成的句子彼此之间相似度很高。

按照以往研究的方法，所有词元级别的指标（即 **BLEU**、**自 BLEU**和**逆向 BLEU**）均计算至 5-gram（de Masson d'Autume 等人，2019年；Ren & Li，2023）。

[5]https://www.tensorflow.org