

---

# GRR-CoCa: LEVERAGING LLM MECHANISMS IN MULTIMODAL MODEL ARCHITECTURES

---

A PREPRINT

 **Jake R. Patock\***

Department of Computer Science  
Rice University  
Houston, TX 77005  
jp157@rice.edu

 **Nicole Catherine Lewis\***

Department of Computer Science  
Rice University  
Houston, TX 77005  
nl66@rice.edu

 **Kevin McCoy**

Department of Statistics  
Rice University  
Houston, TX 77005  
kmm12@rice.edu

**Christina Gomez**

Department of Computer Science  
Rice University  
Houston, TX 77005  
cbg3@rice.edu

**Canling Chen**

Department of Computer Science  
Rice University  
Houston, TX 77005  
cc150@rice.edu

 **Lorenzo Luzi<sup>†</sup>**

Department of Statistics  
Rice University  
Houston, TX 77005  
lorenzo.luzi@rice.edu

July 25, 2025

## ABSTRACT

最先进 (SOTA) 的图像和文本生成模型是多模态模型，与大型语言模型 (LLM) 有许多相似之处。尽管取得了良好的性能，但领先的基础多模态模型架构在架构复杂性上常常落后于当代 LLM。我们提出了 GRR-CoCa，一个改进的 SOTA 对比型描述器 (CoCa) 模型，该模型将高斯误差门控线性单元、均方根归一化和旋转位置嵌入引入到文本解码器和视觉转换器 (ViT) 编码器中。每项架构修改均已证明可提高 LLM 的模型性能，但尚未在 CoCa 中应用。我们将 GRR-CoCa 与基准 CoCa 进行了比对，这是一种拥有相同修改后的文本解码器但保留 CoCa 原始 ViT 编码器的模型。我们使用标准预训练和微调工作流程来对比模型在对比和生成任务中的表现。在预训练数据集和三个不同的微调数据集中，GRR-CoCa 显著优于基准 CoCa。预训练改进为对比损失减少 27.25%，困惑度减少 3.71%，CoCa 损失减少 7.15%。平均微调改进为对比损失减少 13.66%，困惑度减少 5.18%，CoCa 损失减少 5.55%。我们展示了 GRR-CoCa 的修改架构在视觉-语言领域中改进了性能和泛化能力。

**Keywords** Vision Language Model · Image Captioning · Architectural Modifications · Vision Transformers · GRR-CoCa

## 1 介绍

Transformer 在包括 Vaswani et al. [2017] 在内的各个领域的深度学习中取得了无数进展，尤其是在自然语言处理 (NLP) Vaswani et al. [2017], Kenton and Toutanova [2019], Dubey et al. [2024] 方面。因果大语言模型 (LLMs)，包括 Meta 的 Llama 系列，尤其得到改进。这些创新使得现代聊天机器人能够执行复杂任务，如代码补全、详细数据分析、数学问题求解和一般问答 Dubey et al. [2024], Liu et al. [2024], Achiam et al. [2023]。

Transformer 架构已应用于其他模态，例如计算机视觉和音频 Alexey [2020], Radford et al. [2023]。在许多计算机视觉任务中，使用 transformer 作为骨干架构已实现了最新的表现 (SOTA) Alexey [2020], Radford et al. [2023]。多模态模型是 transformer 的扩展，可以通过多种信息流激发，例如文本和图像 Dubey et al. [2024]。这些模型能够有效地复制人类推理，并产生超越以往工作的输出 Yu et al. [2022]。多模态性模型进一步建立在单模态文本解码器或单模态视觉编码器 transformer 的基础上 Dubey et al. [2024]。然而，多模态模型导致复杂的空间、文本和关系依赖性，这对即便是基于 transformer 的深度学习模型来说也是一项困难的任务 Yu

---

\*These authors contributed equally to this work

<sup>†</sup> Author to whom correspondence should be addressed.

et al. [2022]。生成一个能在深层次上捕捉这些复杂依赖关系的机器学习模型对于创建高性能且有用的模型至关重要。

由 Yu et al. [2022] 在 2022 年提出的对比 Captioners (CoCa) 模型是一种当前的 SOTA 模型，其经过微调的 CoCa 视觉编码器在 ImageNet 上的前 1 % 准确率实现了 SOTA 表现。虽然 CoCa 在其视觉编码器和文本解码器中引入了当时新颖的基于变压器的方法，但许多现代的最先进模型，例如 Meta 的 Llama 系列，已经纳入了更新的架构改进 Dubey et al. [2024], Yu et al. [2022]。最近的文献发现，结合现代 LLM 架构能够提高单模语言和视觉变压器模型的性能 Dubey et al. [2024], Chu et al. [2024]。因此，非常有必要更新旧的 SOTA CoCa 模型。

为了解决这个挑战，我们对 CoCa 模型的文本解码器层进行了修改：在前馈层中引入了高斯误差门控线性单元 (GEGLU)，在预归一化位置将层归一化 (LayerNorm) 替换为均方根归一化 (RMSNorms)，并将绝对位置编码替换为旋转位置嵌入 (RoPe) Jiang et al. [2024], Shazeer [2020], Su et al. [2024]。此实现使图像描述模型与当前的 SOTA 单模视觉和 LLM 模型保持一致，从而提升了多模态模型的性能 Dubey et al. [2024], Chu et al. [2024], Team et al. [2024], Adler et al. [2024], Jeevan and Sethi [2022]。

无数现代的最先进视觉 Transformer (ViTs) 也不使用现代的大型语言模型架构。之前的工作表明，使用 RoPe 代替绝对位置编码可以生产出表现更好的 ViT Jeevan and Sethi [2022], Chu et al. [2024]。ViTs 的其他有前景的架构包括 GEGLU 和 RMSNorms，因为它们已被证明可以提升自然语言处理和视觉任务的性能 Shazeer [2020], Zhang and Sennrich [2019], Chu et al. [2024]。基于此，我们提出 GRR-CoCa，这是一种更新的 CoCa 模型，在 CoCa 模型的 ViT 中加入大型语言模型架构，以提升模型生产更多特征丰富的图像潜表示的能力，继而可以被下游模型利用。

总结：

- 我们通过将其文本解码器修改为包括 GEGLU、RMSNorm 和 RoPE，生成了一个基线 CoCa 模型作为对照。这使其与现代 SOTA LLM Transformer 架构 Dubey et al. [2024] 对齐。
- 我们引入了 GRR-CoCa。该模型保留了与 Baseline CoCa 相同的增强文本解码器，同时还结合了配备 GEGLU、RMSNorm 和 RoPE 的视觉编码器。这样确保了跨模态的架构一致性，从而提高了性能。
- 我们在标准的预训练到微调流程中训练了 GRR-CoCa 和基础 CoCa 模型。我们研究了 GRR-CoCa 的架构如何提高模型在预训练数据集和三个不同的微调任务上的拟合能力。
- 我们量化了通过几乎不增加模型参数大小的结构修改，在多模态转换器模型中性能显著提升的程度。

## 2 相关工作

### 2.1 近年来 LLM Transformer 的改进

前面提到的开源 SOTA LLM 模型的架构与“Attention is all you need”中提出的原始仅解码器模型有所不同 Vaswani et al. [2017], Yang et al. [2024], Adler et al. [2024], Dubey et al. [2024], Team et al. [2024]。原始架构采用绝对位置编码、LayerNorms、后规范变换器架构以及单隐藏层前馈网络 Vaswani et al. [2017]。更新的架构包括对原始仅解码器变换器的迭代，旨在提高效率 and 表现力 Dubey et al. [2024], Vaswani et al. [2017]。相比之下，Llama 3.1 使用 RoPe、RMSNorm、预规范变换器架构和前馈网络中的门控线性单元 (GLU) Dubey et al. [2024]。RMSNorm 和预规范架构等修改旨在提高模型效率，而 GLU 和 RoPe 增强了模型的表现力 Zhang and Sennrich [2019], Jiang et al. [2024], Shazeer [2020], Su et al. [2024]。

GLU ViT 是一种 transformer 架构，通过修改前馈层来增强模型的表达能力。在标准的 ViT 中，前馈层由输入层、隐藏层和输出层组成，高斯误差线性单元 (GELU) 的激活函数应用于隐藏层。GLU 的实现修改了前馈结构的前半部分。输入不再是通过单一的线性变换然后跟随偏置和激活函数，而是通过两个并行的线性变换及其附带的偏置。然后，其中一个线性变换通过一个激活函数 (例如 GELU)。之后，采用激活输出与另一个线性变换输出之间的 Hadamard 乘积。这产生的门控输出将传递到最后的线性变换以产生该层的最终输出。以下方程式展示了前馈层的这种修改。

利用这些改进的前馈网络架构，在因果语言建模任务中产生了更好的困惑度评分，并提高了在下游 NLP 任务 (如语义分类 Shazeer [2020]) 上的性能。它在许多 NLP 应用和数据集中的性能持续提升，为在视觉变换器中也包括这些改进提供了有力支持。

#### 2.1.1 均方根误差归一化 (RMSNorm)

层归一化被认为对于稳定任何变压器架构至关重要 Vaswani et al. [2017]。然而，由于其可训练参数对于输入向量的重新中心化和重新缩放不变性的处理，当处理深度神经网络时会导致计算开销。2019 年由 Zhang and Sennrich [2019] 提出的比层归一化更简单的替代方案是 RMSNorm。RMSNorm 修改了方程以去除重新中心化

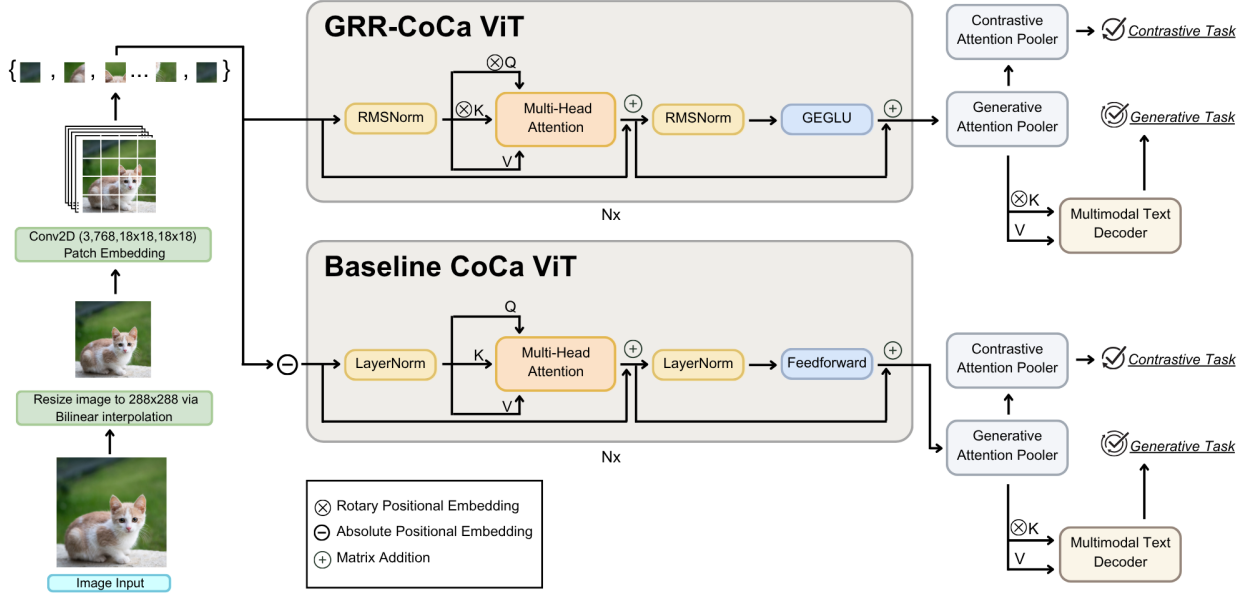


Figure 1: GRR-CoCa 与 Baseline-CoCa 基于 ViT 的视觉编码器架构概述。GRR-CoCa 模型使用 GEGLU、RMSNorm 和 RoPe。Baseline CoCa 模型使用前馈层、LayerNorm 和绝对位置编码。

参数，并消除了计算均值统计的需要，从而降低了该层的计算需求。研究还表明，RMSNorm 在某些任务的评估指标上优于标准的层归一化。更高的效率和经常出现的更优性能支持了我们的假设，即在 GRR-CoCa 中加入 RMSNorm 会提高其在生成式描述任务上的表现。

### 2.1.2 旋转位置嵌入 (RoPe)

Vaswani et al. [2017] 通常在将 token 嵌入传递给 transformer 块之前，将绝对位置编码添加到 token 嵌入中。2021 年，Su et al. [2024] 提出了一种改进原始方法的方案，称为 RoPe。这种方法消除了原有的绝对位置编码，改为应用一组不可训练的线性变换，有效地在一定距离上旋转每个查询和键嵌入，其中旋转的总角度依赖于嵌入在序列中的位置索引。使用 RoPe 在多个自然语言处理任务中（如机器翻译）提高了性能指标。RoPe 可以直接应用于 ViT 的图像补丁嵌入中，并且在 ViT 执行图像分类任务时显示出性能提升。将 RoPE 并入多模态模型的视觉编码器中是一个潜力巨大的尚未深入研究的想法。

提出的将通过之前讨论的 LLM 架构修改而修改的具体 ViT 架构是由 Alexey [2020] 于 2020 年提出的。ViT 采用了传统的原始 transformer 转换，但不同于使用一个可训练的嵌入层的标记输入序列，它将输入图像通过线性形成或卷积操作分割成“补丁”。使用指定窗口大小和步幅的卷积操作，可以将图像划分为补丁，同时将输入通道（通常 RGB 为 3 或者灰度为 1）投射到每个标记的嵌入维度。在发表“图像值得 16x16 个单词” Alexey [2020] 时，ViT 在许多不同的视觉任务中产生了 SOTA 性能，尤其在 ImageNet 图像分类挑战任务上达到了最佳的准确性。

## 2.2 多模态 Transformer

### 2.2.1 CLIP 和 ALIGN

两个值得注意的现代多模态模型是 CLIP Radford et al. [2021] 和 ALIGN Jia et al. [2021]。两者都在大量视觉和文本数据集上训练了大型模型（CLIP 使用视觉转换器，而 ALIGN 使用 EfficientNet 卷积神经网络；两者都使用了文本转换器编码器），以使图像及其配对文本的潜在表示更加相似。所执行的对比损失任务是最大化图像-文本对之间的相似性，并在批处理中最大化不对应的图像-文本对之间的不相似性。这种自监督学习让模型学习将特定图像与其文本表示相关联，反之亦然，结果是在多模态任务中表现更加稳健，同时在单一模态的下游自然语言处理或计算机视觉任务中也表现出色。模型使用文本和视觉信息所学到的高层次抽象被证明是一种强大的学习方式。

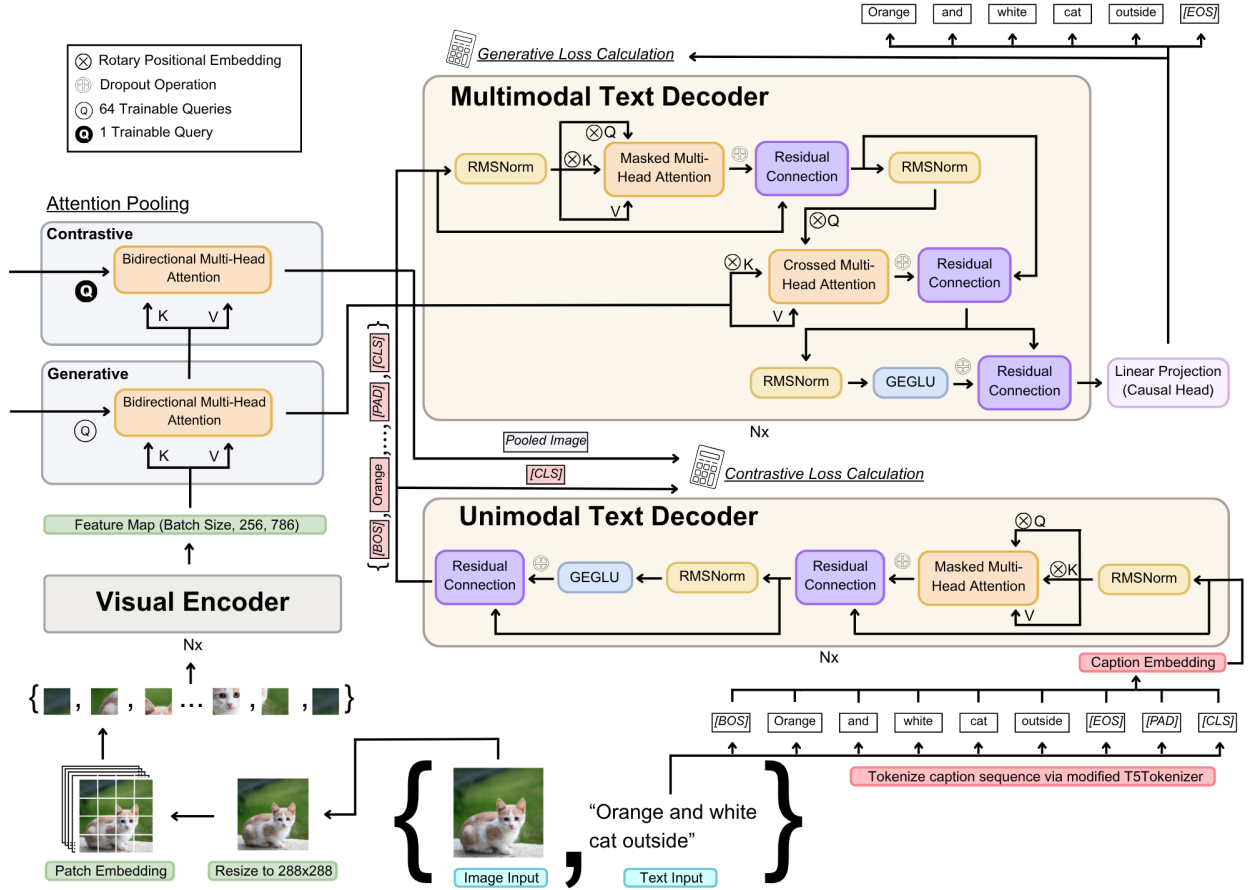


Figure 2: GRR-CoCa 和 Baseline CoCa 的注意力池化器、单模态文本解码器和多模态文本解码器的架构概述。解码器使用 GEGLUs、RMSNorms 和 RoPe。GRR-CoCa 和 Baseline CoCa 使用在 Figure 1 中概述的不同视觉编码器。

### 2.2.2 对比描述器 (CoCa)

CLIP 和 ALIGN 展示了在多模态文本和图像模型中如何通过对比学习在多个基准测试上提供 SOTA 性能 Radford et al. [2019], Jia et al. [2021]。2022 年, Yu et al. [2022] 通过其 CoCa 模型展示了这些对比训练模型的进展 Yu et al. [2022]。CoCa 在所有模态中使用了 transformer 架构 Yu et al. [2022]。它还通过引入另一个文本解码器, 该解码器将文本和图像模型的输出作为输入, 建立在 CLIP 和 ALIGN 简单的对比损失目标之上 Yu et al. [2022]。此外, 一个多模态文本解码器执行字幕损失函数 (简单因果语言建模), 该解码器从单模态解码器获取输出嵌入, 并从视觉编码器获取图像特征图, 进行交叉注意力 Yu et al. [2022]。

这种双重任务的预训练允许模型利用对比自监督学习和生成文本模型训练来生成文本作为输出 Yu et al. [2022]。这有效地创建了一个统一理解的模型, 每个方面 (视觉编码器、单模态文本解码器或多模态文本解码器) 可以单独使用或在多模态任务、NLP 任务或视觉任务中结合使用 Yu et al. [2022]。最初用于训练 CoCa 的数据集是谷歌的专有 JFT-300 M 数据集 Yu et al. [2022]。该数据集包含大约 3 亿对噪声图像-字幕对 Yu et al. [2022]。结果产生了 SOTA 表现, 特别是在 ImageNet 图像分类数据集中使用微调的 CoCa 视觉编码器 Yu et al. [2022]。目前, 这个模型仍然是图像字幕化的 SOTA, 但是文本解码器和图像编码器的架构正在变得过时。

## 3 方法与实验

用于预训练的数据集是概念性标题 1200 万 (CC12M) 数据集。CC12M 被用来模拟原始 CoCa 模型使用的更大的专有 JFT 图像标题数据集。CC12M 包含一般对象的图像标题对, 例如 “维多利亚风格的餐厅”。我们随

机将 CC12M 分为 10,445,110 个图像标题对用于训练集, 549,743 个图像标题对用于验证集 (95% 训练, 5% 验证)。

### 3.0.1 微软常用物体数据集 (MSCOCO)

第一个微调数据集是微软的 MSCOCO (Microsoft Common Objects in Context) 数据集 Lin et al. [2014]。MSCOCO 的样本与 CC12M 的分布相似, 因为它们都包含常见的物体。MSCOCO 由原作者预先划分为训练和验证集 Lin et al. [2014]。每个训练实例通常有五个不同的描述, 因此训练集中的图像和描述对总数为 591,753, 验证集包含 25,014 个图像和描述对。

第二个微调数据集是 Radiology Objects in COntext (ROCO) Version 2 数据集, 这是一个更复杂的医学图像描述数据集。ROCO 用于模拟迁移学习场景, 因为它与 CC12M 的分布相差较远。在诸如 CLIP 和 PubMedClip 等多模态模型中, 先在一般知识上进行预训练然后在特定医学数据集上微调已被发现是一种实用的方法。ROCO 已被预先分为训练集和验证集, 训练集包含 69,866 个图像描述对, 验证集包含 9,904 个图像描述对。

最后, GRR-CoCa 和基准 CoCa 模型在 Flickr 30K 数据集上进行了微调 Young et al. [2014]。这个常用的基准数据集包含大约 30k 张图片, 每张图片都有五个人工标注的标题 Young et al. [2014]。数据集使用的训练-验证划分是 Karpathy 划分, 其中包含 29,000 张训练图片和 1,000 张验证图片 Karpathy and Fei-Fei [2015]。

为了将每个数据集的标题处理为输入 ID, 我们使用了一个修改过的 T5 基础分词器 Raffel et al. [2020]。我们在词汇表中添加了“序列开始”标记 (BOS) 和分类标记 (CLS)。然后, 将 BOS 标记添加到所有标题的开头, 并在标题中最后一个非填充标记后面添加“序列结束”标记 (EOS)。原生 T5 PAD 标记用于将每个标题填充到统一长度。在 PAD 标记之后附加 CLS 标记。最后, 在 CLS 标记之后附加一个单独的 PAD 标记, 以生成正确形式的因果语言建模。这一过程对所有数据集中的所有标题进行。

所有图像都使用双线性插值调整为  $288 \times 288$  像素, 转换为 RGB, 并缩放到  $[0, 1]$  区间。最后, 这些值使用标准的 ImageNet 通道均值和标准差 Deng et al. [2009] 进行归一化。

### 3.1 模型架构

我们构建了两种不同的多模态 CoCa 模型变体。这两种 CoCa 模型变体在其文本解码子模型 (即, 单模态和多模态解码器) 中使用了现代 LLM 架构 (GEGLUs、RoPe 和 RMSNorms), 如 Figure 2 所示。这确保了结果的任何改进都是由于视觉编码器的修改。“Baseline CoCa”使用了一个 ViT 子模型, 其架构遵循自 Alexey [2020] 的原始架构。GRR-CoCa 使用了一个修改的 ViT, 用 GEGLU 替换了前馈网络, 用 RoPe 替换了绝对位置编码, 并用 RMSNorms 替换了 LayerNorms Su et al. [2024], Shazeer [2020], Zhang and Sennrich [2019]。表 1 展示了 GRR-CoCa 和 Baseline CoCa 使用的视觉编码器。

所使用的池化与原始 CoCa 论文 Yu et al. [2022] 中描述的注意力池化器相同。这些是自注意力层, 其中键和值来源于视觉编码器输出和  $n$  查询 Yu et al. [2022]。参数  $n$  确定我们正在训练的查询数量, 并且与从此注意力块输出的嵌入数量相同 Yu et al. [2022]。对于 GRR-CoCa, 我们首先将视觉编码器输出 (256 图像补丁嵌入) 发送到生成池化器, 该池化器将 256 补丁投影到模型的上下文长度 64 Yu et al. [2022]。然后, 此输出作为交叉注意力中的上下文传递 Yu et al. [2022]。同时, 这个生成输出池化器被传递到另一个对比池化器, 将这些 64 嵌入投影到用于对比损失函数的单一嵌入, 如 Figure 2 Yu et al. [2022] 中所示。这个级联池化器结构被认为是 Yu et al. [2022] 测试过的最佳结构。

模型的超参数选择类似于原始 CoCa 模型 Yu et al. [2022]。视觉编码器、单模态解码器和多模态解码器的嵌入维度都是 768。每个编码器/解码器子模型都有 12 个 Transformer 块。子模型的注意力层各有 12 个注意力头。在预训练期间应用了一个通用的 0.15 的 dropout, 而在微调时应用了 0.1 的 dropout。非 GEGLU 前馈网络中隐藏层的大小计算为嵌入维度乘以 4。由于在 GEGLU 中添加额外的线性变换会导致模型整体参数数量增加, 如果在前馈网络中对隐藏层大小使用相同的  $4 \times$  缩放因子, 我们反而通过 2.7 因子缩放 GEGLU 中的隐藏层大小, 以保持模型的参数大小相似 Shazeer [2020]。模型超参数的摘要见表 1。GRR-CoCa 比基线 CoCa (393,863,270) 多了 0.17% 个可训练参数 (394,548,926)。模型的上下文长度设置为 64 个 token, 带有添加的 token 的 T5 分词器使词汇量达到了 32,102。

### 3.2 损失函数

Baseline CoCa 和 GRR-CoCa 同时在对标和生成性字幕任务上训练。对标任务涉及训练模型, 使得它们的图像潜在表示和文本潜在表示具有相似的嵌入。

对比损失函数通过减小图像-文本对在嵌入空间中的距离, 同时增加不匹配对的距离 Yu et al. [2022]。这有效地推动模型将图像的潜在表示与文本的潜在表示关联起来。较低的对比损失值表明更强的对齐程度, 反映了架构改进的影响。对于每个图像-文本对的批次, 潜在表示通过对比损失进行评估, 如下所示 Yu et al. [2022] :

Table 1: GRR-CoCa 和 Baseline CoCa 中使用的模型架构和分词器超参数。

| Architecture Hyperparameter           | Value  |
|---------------------------------------|--------|
| Embedding Dimension                   | 768    |
| Number of Attention Heads per Model   | 12     |
| Number of Blocks per SubModel         | 12     |
| Non-GEGLU Feedforward Hidden Scaler   | 4      |
| GEGLU Feedforward Hidden Layer Scaler | 2.7    |
| Token Context Length                  | 64     |
| Vocabulary Size                       | 32,102 |
| Dropout During Pretraining            | 0.15   |
| Dropout During Fine-tuning            | 0.1    |

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N} \left[ \sum_{i=1}^N \log \left( \frac{\exp(x_i^\top y_i / \sigma)}{\sum_{j=1}^N \exp(x_i^\top y_j / \sigma)} \right) + \sum_{i=1}^N \log \left( \frac{\exp(y_i^\top x_i / \sigma)}{\sum_{j=1}^N \exp(y_i^\top x_j / \sigma)} \right) \right] \quad (1)$$

其中

生成式描述任务使用自回归因果语言模型进行训练，在该模型中，模型学习在给定所有前面词语的情况下，最大化每个词语的概率 Yu et al. [2022]。这个损失的公式为

其中

- $T$  是目标序列中的标记总数，
- $y_t$  是我们试图在索引  $t$  处预测的标记，
- $Y_{<t}$  是序列中在索引  $t$  之前的所有标记，
- $x$  是我们要为其生成标题的输入图像，并且
- $P_\theta$  是具有参数  $\theta$  的模型预测  $y_t$  的概率。

在实际应用中，交叉熵损失函数（softmax 操作结合负对数似然损失）被应用于模型的 logits。因果建模中忽略了三个 tokens，分别是 BOS、PAD 和 CLS tokens。考虑到这一点，得出下面的最终生成式字幕损失函数。

$$\mathcal{L}_{\text{cap}} = -\sum_{t=1}^T y_t \cdot \log \left( \frac{\exp(\hat{y}_t)}{\sum_{t=1}^T \exp(\hat{y}_t)} \right) \cdot 1_{\{y_t \neq \text{ignore\_index}\}} \quad (2)$$

- $T$  是目标序列中的标记总数，
- $y_t$  是我们在索引  $t$  处试图预测的标签向量的元素，并且
- $\hat{y}$  是我们试图在索引  $t$  处预测的 logits 向量的元素。

所使用的生成性字幕评估指标是困惑度，其计算方法是每个标记的平均交叉熵损失的指数形式：

$$\text{perplexity} = \exp \left( \frac{\mathcal{L}_{\text{cap}}}{T} \right) \quad (3)$$

其中  $\mathcal{L}_{\text{cap}}$  是所有有效标记的生成字幕损失之和， $T$  是序列中有效（非忽略）标记的数量。

然后，每个损失项都由超参数 ( $\lambda_{\text{Con}}$  和  $\lambda_{\text{Cap}}$ ) 加权，并相加生成 CoCa 损失函数，Yu et al. [2022] 在训练过程中同时最小化两个任务。

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}} * \mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}} * \mathcal{L}_{\text{Cap}} \quad (4)$$

在预训练期间，lambda 的权重被设置为  $\lambda_{\text{Con}} = 2$  和  $\lambda_{\text{Cap}} = 1$ ，而在微调期间被设置为  $\lambda_{\text{Con}} = 1$  和  $\lambda_{\text{Cap}} = 2$ 。这确保预训练专注于对比损失的最小化，并且生成式字幕损失不会过度拟合预训练数据集，从而允许模型灵活地学习特定微调数据集的字幕。在微调过程中，模型的重要功能是其产生的字幕。因此，生成式字幕损失系数被加倍，以确保与模型目标更紧密地对齐。

Table 2: 在预训练 (CC12M) 和微调 (MSCOCO、ROCO 和 Flickr30K) 数据集上, 基线 CoCa 和 GRR-CoCa 的验证评估指标。GRR-CoCa 在所有数据集上的所有评估指标上表现更好。

| Metric                                | Baseline CoCa | GRR-CoCa | % Change |
|---------------------------------------|---------------|----------|----------|
| Pretraining: CC12M Validation Set     |               |          |          |
| CoCa Loss                             | 3.2864        | 3.0516   | -7.15 %  |
| Perplexity                            | 12.9976       | 12.5151  | -3.71 %  |
| Contrastive Loss                      | 0.3610        | 0.2626   | -27.25 % |
| Fine-Tuning: MSCOCO Validation Set    |               |          |          |
| CoCa Loss                             | 4.6955        | 4.4090   | -6.10 %  |
| Perplexity                            | 5.4669        | 5.1523   | -5.75 %  |
| Contrastive Loss                      | 1.2971        | 1.1296   | -12.92 % |
| Fine-Tuning: ROCO Validation Set      |               |          |          |
| CoCa Loss                             | 7.0609        | 6.8708   | -2.69 %  |
| Perplexity                            | 12.0258       | 11.7377  | -2.40 %  |
| Contrastive Loss                      | 2.0900        | 1.9479   | -6.80 %  |
| Fine-Tuning: Flickr30K Validation Set |               |          |          |
| CoCa Loss                             | 5.4383        | 5.0111   | -7.86 %  |
| Perplexity                            | 7.9568        | 7.3686   | -7.39 %  |
| Contrastive Loss                      | 1.2908        | 1.0164   | -21.25 % |

### 3.3 训练技巧

用于模型生成的框架是 PyTorch, 计算使用了四个 Tesla L40S GPU Paszke et al. [2019]。每个 GPU 的批量大小为 48, 加上 4 步梯度积累, 使得每步总模拟批量大小为 768 样本 ( $48 \times 4 \times 4 = 768$ )。使用的优化器是现代的 AdamW, 默认 betas 为 (0.9 和 0.999), 权重衰减为 0.001 Loshchilov and Hutter [2017]。使用的学习率调度方案是简单的线性预热以保持早期稳定性, 持续 2000 步, 然后是强大的余弦退火带有暖重启调度器 (CAWR) Loshchilov and Hutter [2016], Liu et al. [2024]。线性预热仅用于预训练。在预训练期间, CAWR 的最大起始学习率为  $1e-4$ , 最小起始学习率为  $1e-6$ 。在微调期间, CAWR 的最大和最小学习率分别为  $1e-5$  和  $1e-7$ 。CAWR 的循环长度为每个训练周期 1 学习率循环。使用梯度裁剪将最大范数限制为 1 以保持稳定性。

为了在微调数据集上训练到收敛, 使用了验证损失监控的提前停止器。耐心设定为微调数据集上的 9 个 epoch。然而, 为了确保尽可能接近地拟合验证集, 提前停止器中使用了一个软重置机制。这一技术会将模型重置到验证损失表现最佳的状态, 如果损失在最近的 3 个 epoch 中没有改善的话。它还将 CAWR 的最大和最小学习率降低一个因子 100, 并从重置状态开始重新训练。这使得模型能够在达到局部最小值后动态降低学习率, 并由于参数搜索空间的地理位置学习率过大而开始交替。实质上, 使用这种“学习率降低与软重置”技术, 使得模型在三个不同的最大和最小学习率区间内尝试降低验证损失, 然后终止训练。

## 4 结果

GRR-CoCa 和基线 CoCa 模型都在 CC12M 数据集上进行了 21 个 epoch (大约 30 万步) 的预训练。GRR-CoCa 模型在验证分区上所有评估指标 (CoCa 损失、困惑度和对比损失) 上都优于基线 CoCa 模型。

在预训练之后, Baseline CoCa 和 GRR-CoCa 都在 MSCOCO、ROCO 和 Flickr30K 数据集上进行了收敛训练。“线性预热”和“带软重置的学习率下降”调度器改善了训练时间, 并确保从 GRR-CoCa 和 Baseline CoCa 中获得最大性能。GRR-CoCa 在所有微调数据集上产生了更好的评估指标, 遵循了预训练结果的趋势。相对于基线的评估指标和改进显示在 Table 2 中。

在 CC12M 上, GRR-CoCa 的 ViT 架构展现出了更高的能力来学习训练数据并泛化到未见过的验证数据, 使用的参数大小、训练周期数和超参数配置与基准 CoCa 大致相同。

在预训练中改进最多的评估指标是对比损失。注意, 由于 lambda 的选择 ( $\lambda_{\text{Cap}} = 1$ ,  $\lambda_{\text{Con}} = 2$ ), 生成式字幕任务的损失权重仅为对比损失的一半。GRR-CoCa 在几乎没有增加模型大小的情况下, 大幅提高了模型实现更低对比损失的能力。这可能是由于对比任务对于网络建模来说比捕捉自然语言和图像之间复杂的相互依赖关系要容易, 如同在生成任务中所做的一样。困惑度也有所改善, 但不那么明显。因此, 对 ViT 的修改在

CoCa 模型中改善了对比和生成字幕任务。这一趋势在微调数据集中也得到了验证，其中  $\lambda_{\text{Cap}} = 2$ ， $\lambda_{\text{Con}} = 1$ 。GRR-CoCa 在所有评估指标上都被证明优于基线 CoCa。

## 5 讨论

在对比和生成字幕任务中 GRR-CoCa 所取得的改进表明，将 GEGLUs、RoPe 和 RMSNorms 引入 ViT 使其比在 Baseline CoCa 中使用的原始 ViT 架构产生了更具特征丰富性的图像潜在表示。这些结果表明，GRR-CoCa 比 Baseline CoCa 学习得更快更深入。

我们显示了在各种任务和模式中，RoPe 倾向于提供比绝对位置编码更好的位置编码。我们具体的 RoPe 实现是在每个 ViT 块中注入位置编码。这允许模型通过在每个块中重新注入位置信息，将补丁嵌入的位置信息保留到更深的层中。一个可能的解释是，随着嵌入经历后续块，绝对位置编码信息在更深的 Transformer 层中减弱，而 RoPe 保持了保留信息的能力，从而提高了其有效性。

GEGLUs 带来的改进比较模糊，但主要归因于它们所创建的信息瓶颈。GEGLUs 限制某些信息通过它们，有效地放大或缩小在每个嵌入中的这些信息。我们推测这种属性可以让每个嵌入根据注意力层的结果进行更细致的操控。信息过滤器可能有助于保留每个嵌入中更具普遍性的信息，并过滤掉嵌入中的噪声，从而实现更快速的学习和更好的泛化能力。

RMSNorm 对 LayerNorm 的简化仅仅是消除了重新中心化参数。性能的提高表明，这一层中的关键参数是来自 LayerNorm 的  $\gamma$  缩放参数，而不是  $\beta$  参数。移除  $\beta$  参数会使训练数据中的噪声建模减少，从而改善总体的泛化能力。

## 6 结论

### 6.1 影响

多模态模型不仅在视觉-语言配对任务中表现出广泛的实用性，而且作为单模态流水线中的组件，其预训练的视觉编码器或文本解码器可以为专业应用进行微调 Yu et al. [2022], Radford et al. [2021]。在这项工作中，我们展示了对 ViT 的结构增强，特别是 GEGLUs、RoPe 和 RMSNorms，在不增加模型规模的情况下大幅提高了鲁棒性和下游性能。与基线的 CoCa 相比，我们的 GRR-CoCa 在一次预训练和三个多样化微调数据集上，分别对 CoCa 损失、困惑度和对比损失实现了平均 5.95 %、4.81 % 和 17.05 % 的改善。这些修改在对比目标和生成式标注任务上带来了持续的增益，并加深了模型的学习表征。通过展示 GEGLUs、RoPe 和 RMSNorms 可以无缝集成到任何基于 ViT 的视觉编码器中，并实证量化其影响，我们为设计下一代基础模型提供了实用指导。采用这些技术可实现更快的训练、更低的损失、更小的高性能模型，以及在广泛的下游应用中提高的准确率。

### 6.2 未来工作

本研究的结果和见解可以为未来的工作提供几个有前途的方向。首先，未来的研究应该调查模型大小和数据集复杂性与修改后的文本解码器和视觉编码器修改的缩放关系。然后，GRR-CoCa 可以重新在 JFT 数据集上进行训练，并在原始 CoCa 论文中展示的任务上进行基准测试 Yu et al. [2022]。这将提供对这些修改如何影响多模态模型性能的全面审查。选择更大的基础数据集，如 LAION-5B 或 JFT 3B，与更大的 CoCa 模型 (2B 参数) 配对，也将为我们当前的发现提供更多支持 Schuhmann et al. [2022], Yu et al. [2022]。

最后，除了简单的扩展，未来的工作还应探索这些修改对其他视觉任务的适应性，例如视觉问答、视觉涵义推理或跨模态检索。这将有助于确定在图像字幕生成中观察到的改进是否能够推广到其他需要对图像和文本关系进行细致理解的任务。

## References

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.
- Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified LLaMA interface for vision tasks. *arXiv preprint arXiv:2403.00522*, 2024.
- Zixuan Jiang, Jiaqi Gu, Hanqing Zhu, and David Pan. Pre-RMSNorm and Pre-CRMSNorm transformers: equivalent and efficient Pre-LN transformers. *Advances in Neural Information Processing Systems*, 36, 2024.
- Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. Nemotron-4 340B technical report. *arXiv preprint arXiv:2406.11704*, 2024.
- Pranav Jeevan and Amit Sethi. Resource-efficient hybrid X-formers for vision. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2982–2990, 2022.
- Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. URL <https://cocodataset.org/#home>.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

- Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.