IS ATTENTION BETTER THAN MATRIX DECOMPOSITION?

注意力比矩阵分解好吗？

摘要

作为现代深度学习的一个基本要素，注意力机制，尤其是自我注意力，在全局关联发现中起着至关重要的作用。然而，在对全局背景进行建模时，手工制作的注意力是否不可替代？我们有趣的发现是，在编码长距离依赖关系的性能和计算成本方面，自我注意并不比20年前开发的矩阵分解（MD）模型更好。我们将全局上下文问题建模为一个低秩恢复问题，并表明其优化算法可以帮助设计全局信息块。然后，本文提出了一系列的Hamburgers，在这些Hamburgers中，我们采用解决MD的优化算法，将输入表征分解为子矩阵，并重建低秩嵌入。当仔细处理通过MDs回传的梯度时，具有不同MDs的Hamburgers可以对流行的全局上下文模块self-attention有良好的表现。在学习全局背景至关重要的视觉任务中进行了全面的实验，包括语义分割和图像生成，证明了比self-attention及其变体的明显改进。代码可用。

代码链接：<https://github.com/Gsunshine/Enjoy-Hamburger>

1 引言

自从自我注意和转化器（Vaswani等人，2017）在捕捉长距离依赖性方面显示出比递归神经网络和卷积神经网络更明显的优势后，注意力被计算机视觉（Wang等人，2018；Zhang等人，2019a）和自然语言处理（Devlin等人，2019）广泛采用于全局信息挖掘。然而，在对全局环境进行建模时，手工制作的注意力是不可替代的吗？本文重点介绍了一种设计全局语境模块的新方法。其关键思想是，如果我们将全局语境这样的归纳偏向表述为一个目标函数，那么使目标函数最小化的优化算法可以构建一个计算图，也就是我们在网络中需要的架构。我们通过为最具代表性的全局语境模块--自我注意--开发一个对应物来具体说明这个想法。考虑到在网络中提取全局信息是寻找字典和相应的代码来捕捉内在的相关性，我们将上下文发现建模为输入张量的低秩恢复，并通过矩阵分解来解决它。然后，本文提出了一个全局相关块，Hamburger，通过采用矩阵分解将学习到的表征分解成子矩阵，从而恢复干净的低秩信号子空间。解决矩阵分解的迭代优化算法定义了中心计算图，即Hamburger的架构。我们的工作利用了作为Hamburger基础的矩阵分解模型，包括矢量量化（VQ）（Gray & Neuhoff，1998）、概念分解（CD）（Dhillon & Modha，2001）和非负矩阵分解（NMF）（Lee & Seung，1999）。此外，我们没有直接应用反向传播算法（BPTT）（Werbos等，1990）来区分迭代优化，而是采用截断的BPTT算法，即一步梯度，来有效地反向传播梯度。我们在全局背景已被证明至关重要的基本视觉任务中说明了Hamburger的优势，包括语义分割和图像生成。实验证明，当避免不稳定的梯度通过MD的迭代计算图回传时，优化设计的Hamburger可以与最先进的注意力模型竞争。Hamburger在PASCAL VOC数据集（Everingham等人，2010）和PASCAL Context数据集（Mottaghi等人，2014）的语义分割中创造了新的最先进的记录，并在ImageNet（Deng等人，2009）的大规模图像生成中超越了现有的GANs的注意力模块。本文的贡献列举如下。

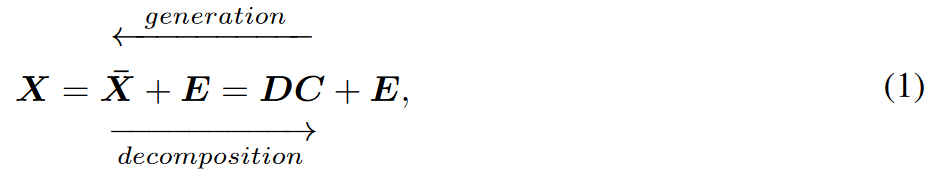
- 我们展示了一种设计全局信息块的白盒方法，即通过将最小化目标函数的优化算法，其中对全局相关性的建模被表述为一个低秩恢复问题，变成了架构。

- 我们提出了Hamburger，一个具有O(n)复杂度的轻便而强大的全局上下文模块，在语义分割和图像生成上超过了各种注意力模块。

- 我们发现，在网络中应用MD的主要障碍是通过其迭代优化算法产生的不稳定的后向梯度。作为一个实际的解决方案，所提出的单步梯度有利于用MD训练Hamburger。  
  
2 方法

2.1 热身

由于矩阵分解对提议的Hamburger至关重要，我们首先回顾矩阵分解的概念。一个常见的观点是，矩阵分解将观察到的矩阵分解为几个子矩阵的乘积，例如，单值分解。然而，一个更有启发性的观点是，通过假设生成过程，矩阵分解作为生成的逆过程，将构成复杂数据的原子分解。从原始矩阵的重建中，矩阵分解恢复了观察数据的潜在结构。假设给定的数据被排列成一个大矩阵X = [x1, - -, xn] ∈ Rd×n的列。一个一般的假设是，在X中隐藏着一个低维的子空间，或者多个子空间的联合。也就是说，存在一个字典矩阵D=[d1, - -, dr]∈Rd×r和相应的编码C=[c1, - -, cn]∈Rr×n，X可以表示为



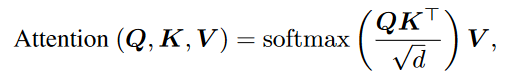
其中，̄X∈Rd×n是输出的低秩重建，E∈Rd×n是要丢弃的噪声矩阵。这里我们假设恢复的矩阵̄X具有低秩属性，即



通过假设矩阵D、C和E的结构，可以得出不同的MD（Kolda & Bader，2009；Udell等人，2016）。MD通常被表述为带有各种约束条件的目标，然后通过优化算法来解决，其经典应用包括图像去噪（Wright等人，2009；Lu等人，2014）、画中画（Mairal等人，2010）和特征提取（Zhang等人，2012）。

2.2 提议的方法

我们专注于为网络建立全局性的上下文模块，而不需要费力地手工设计。在开始讨论之前，我们先回顾一下有代表性的手工设计的语境模块自我注意。注意力机制旨在从大量无意识的上下文中找到一组概念进行进一步的有意识推理（Xu等人，2015；Bengio，2017；Goyal等人，2019）。作为一个代表，自我注意（Vaswani等人，2017）被提出用于学习机器翻译中的长程依赖。



其中，Q、K、V∈Rn×d是通过线性变换从输入中投影出来的特征。自我关注通过一次关注所有的标记来提取全局信息，而不是典型的递归神经网络的逐一处理。

尽管自我注意及其变体取得了巨大的成功，但研究人员面临着（1）在自我注意的基础上开发新的全局情境模块，通常是通过手工制作的工程，以及（2）解释为什么目前的注意模型能够发挥作用。本文绕过了这两个问题，找到了一种通过定义明确的白盒工具箱轻松设计全局情境模块的方法。我们尝试将人类的归纳偏见，如全局语境，表述为一个目标函数，并使用优化算法来解决这样的问题，以设计模块的架构。优化算法创建一个计算图，接受一些输入，并最终输出解决方案。我们将优化算法的计算图应用于我们情境模块的中心部分。基于这种方法，我们需要将网络的全局语境问题建模为一个优化问题。以卷积神经网络（CNN）为例进行进一步讨论。在我们输入一幅图像后，网络会输出一个张量X∈RC×H×W。由于张量可以被看作是一组HW C维的超像素，我们将张量展开为一个矩阵X∈RC×HW。当模块学习长距离依赖关系或全局背景时，隐藏的假设是超像素是固有的相关的。为了简单起见，我们假设超像素是线性依赖的，这意味着X中的每个超像素可以表示为基数的线性组合，其元素通常远小于HW。在理想情况下，隐藏在X中的全局信息可以是低等级的。然而，由于香草CNN对全局环境的建模能力较差（Wang等人，2018；Zhang等人，2019a），学到的X通常会被冗余信息或不完整信息所破坏。上述分析提出了一种潜在的方法来为全局上下文建模，即通过完成展开的矩阵X中的低秩部分̄ X并丢弃噪声部分E，使用公式（1）中描述的经典矩阵分解模型，同时过滤掉冗余和不完整性。因此，我们将学习全局背景建模为一个以矩阵分解为解决方案的低秩恢复问题。使用第2.1节的概念，矩阵分解的一般目标函数为

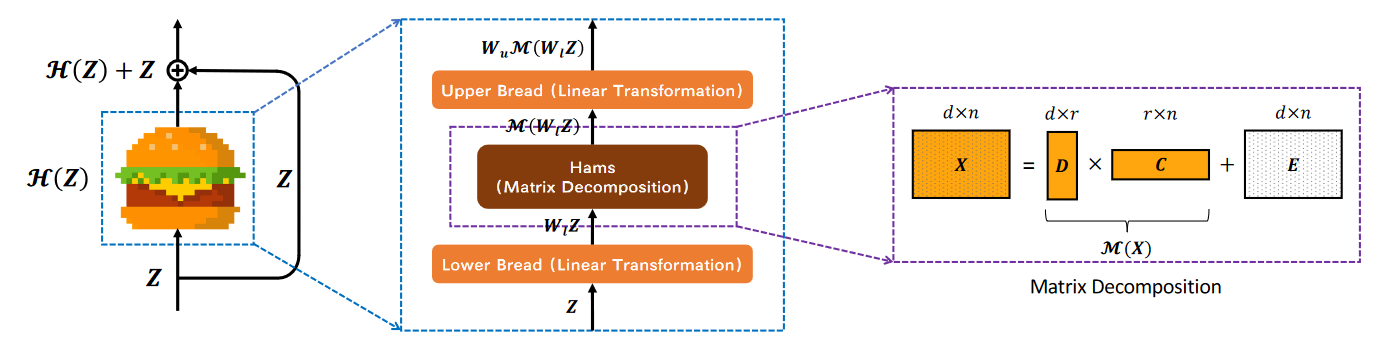
  
其中L是重建损失，R1和R2是字典D和代码C的正则化项。将最小化公式（4）的优化算法称为M。为了帮助读者进一步理解这种建模，我们还在附录G中提供了一个更直观的说明。在后面的章节中，我们将介绍我们的全局上下文模块Hamburger，然后讨论M的详细MD模型和优化算法。最后，我们通过矩阵分解处理反向传播的梯度问题。

2.2.1 hamburger

堡包由一片 "火腿"（矩阵分解）和两片 "面包"（线性变换）组成。顾名思义，Hamburger首先用线性变换Wl∈Rd×dz将输入Z∈Rdz×n映射到特征空间，即 "下面包"，然后用矩阵分解M解决一个低秩信号子空间，对应于 "火腿"，最后用另一个线性变换Wu∈Rdz×d将提取的信号变换到输出，称为 "上面包"。



其中M是矩阵分解，以恢复清晰的潜伏结构，作为全局非线性的功能。2.2.2节中讨论了M的详细架构，即对X进行因子化的优化算法。图1描述了Hamburger的架构，



它通过批量归一化（BN）（Ioffe & Szegedy, 2015）、跳过连接与网络协作，最后输出Y 。



2.2.2 HAMS

本节介绍 "火腿 "的结构，即公式（5）中的M。如上节所述，通过将全局信息发现表述为MD的一个优化问题，解决MD的算法自然而然地组成了M。 M将 "下面包 "的输出作为其输入，并计算出一个低秩重建作为其输出，分别表示为X和̄X

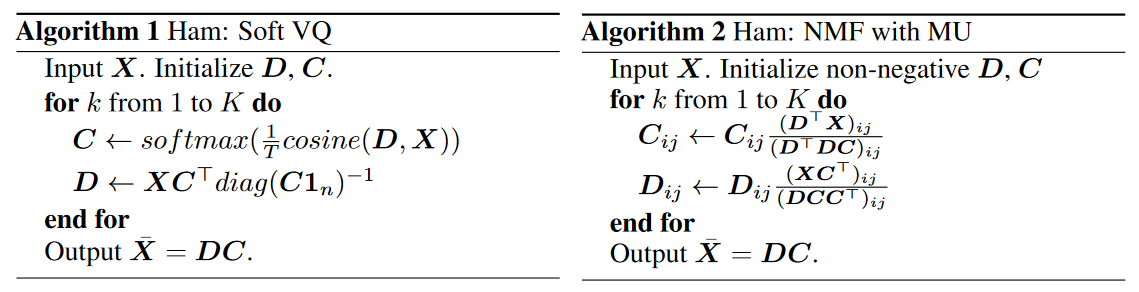


我们研究了M的两个MD模型，即矢量量化（VQ）和非负矩阵分解（NMF）来解决D和C并重建̄ X，而将概念分解（CD）留给附录B。对所选的MD模型进行了简单介绍，因为我们努力说明低秩归纳偏置和优化驱动设计方法对全局上下文模块的重要性，而不是任何具体的MD模型。我们更倾向于将MD部分作为一个整体来抽象，即本文中的M，并关注Hamburger如何在整体上显示其优越性。

**矢量量化**矢量量化（VQ）（Gray & Neuhoff, 1998）是一种经典的数据压缩算法，可以被表述为矩阵分解方面的一个优化问题：。



其中，ei是典型的基础向量，ei = [0, - -, 1, - -, 0]> ith。最小化公式（8）中目标的解决方案是K-means（Gray & Neuhoff, 1998）。然而，为了确保VQ是可微分的，我们用sof tmax和余弦相似度取代硬性的arg min和欧氏距离，从而得出Alg. 1，其中cosine(D, X)是一个相似性矩阵，其条目满足，sof tmax是逐列应用，T是温度。此外，当T→0时，我们可以通过一个单热矢量获得一个硬分配。



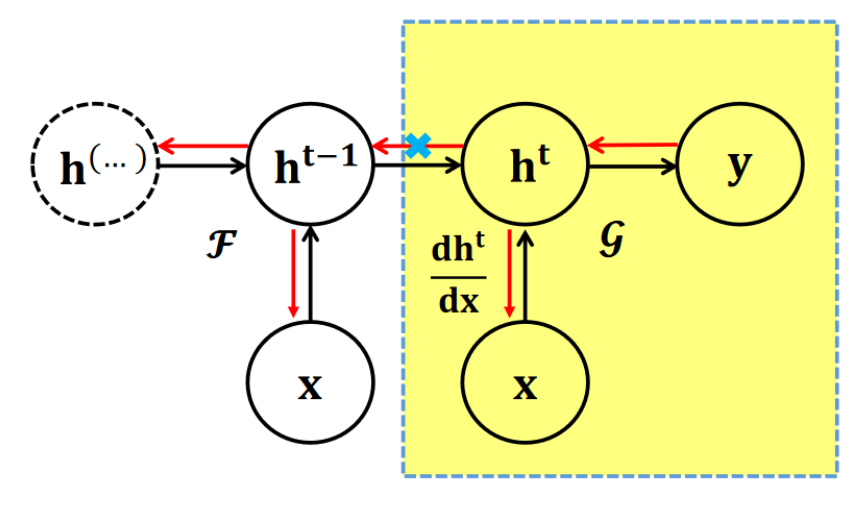
**非负矩阵分解** 如果我们对字典D和编码C施加非负约束，就会导致非负矩阵分解（NMF）（Lee & Seung, 1999）。



为了满足非负约束，我们在将X放入NMF之前添加一个ReLU非线性。我们应用Alg.2中的乘法更新（MU）规则（Lee & Seung, 2001）来解决NMF，这保证了收敛性。作为白盒全局模块，VQ、CD和NMF都是简单明了的，显示了显著的效率。它们被制定为优化算法，主要由复杂度为O(ndr)的矩阵乘法组成，比r n时自我关注的复杂度O(n2d)要低得多。这三个MD都是内存友好的，因为它们避免生成一个大的n×n矩阵作为中间变量，如公式（3）中自我关注的Q和K之积。在后面的章节中，我们的实验证明MD至少与自力更生相当，尽管M的架构是通过优化创建的，看起来与经典的点乘自力更生不同。

2.3 单步梯度

由于M涉及一个优化算法作为其计算图，将其融合到网络中的一个关键是迭代算法如何回传梯度。类似于RNN的优化行为表明反向传播（BPTT）算法（Werbos等人，1990）是区分迭代过程的标准选择。下面我们首先回顾一下BPTT算法。然而，在实践中，BPTT的不稳定梯度确实损害了Hamburger的性能。因此，我们建立一个抽象的模型来分析BPTT的缺点，并试图找到一个务实的解决方案，同时考虑到MD作为一个优化算法的性质。如图2所示，



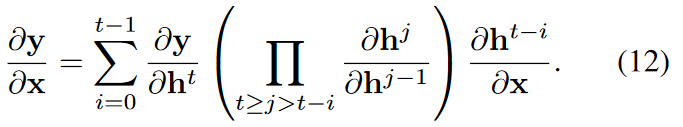
x、y和hi分别表示时间步骤i的输入、输出和中间结果，而F和G是运算符。在每个时间步骤中，模型收到由底层网络处理的相同输入x。



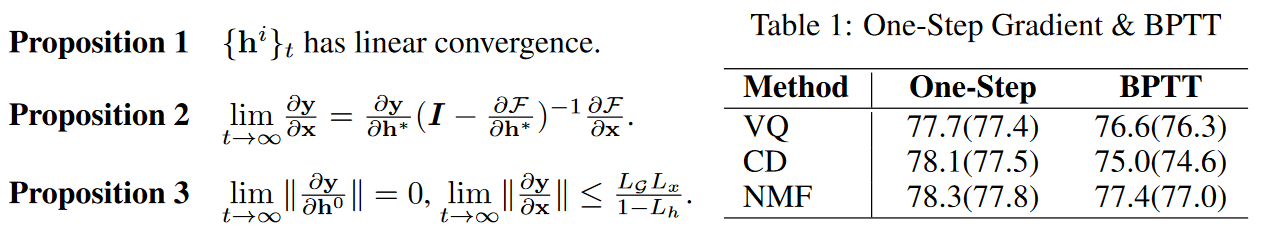
中间的结果hi都被丢弃了。只有最后一步的输出ht，通过G进行输出y。



在BPPT算法中，根据Chain规则，从输出y到输入x的Jacobian矩阵已经给出。

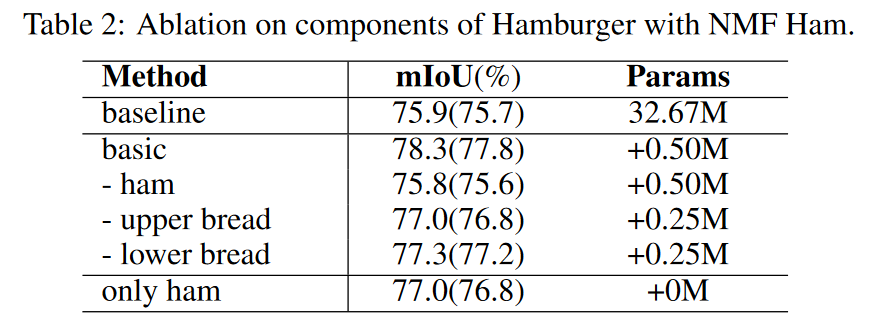


一个思想实验是考虑t→∞，导致一个完全收敛的结果h∗和公式（12）中的无限项。我们假设F和G都是Lipschitz，常数为Lh（相对于h），Lx（相对于x），以及LG，并且Lh<1。请注意，这些假设适用于大量的优化或数值方法。那么我们有



此外，当∂F ∂h的最大特征值，即F的Lipschitz常数接近1时，Jacobian矩阵∂y ∂x受到条件不良项（I - ∂F ∂h∗ ）-1的影响。的Lipschitz常数接近1，而其最小特征值通常保持在0附近，因此限制了梯度在参数空间中搜索广义解的能力。通过优化算法，梯度的规模和频谱不稳定，表明将BPTT直接应用于Hamburger是不可行的，表1中的实验证实了这一点。1，使用与第3.1节相同的消融设置。这个分析给了我们一个可能的解决方案。需要注意的是，在BPTT算法中存在多个雅各布矩阵的乘法∂hj ∂hj-1和一个无限级数的求和，导致梯度的不可控性。这启示我们，在保留梯度的主导项的同时，放弃一些次要项，以确保方向大致正确。考虑到公式（12）的条款是一个系列，即{ ∂y ∂ht ( ∏ t≥j>t-i ∂hj ∂hj -1 ) ∂ht-i ∂x }i，如果其条款的规模以指数方式衰减，用这个系列的第一个条款来近似梯度是有意义的。梯度的第一项来自优化的最后一步，导致单步梯度的出现。  


根据命题2，当t→∞时，一步梯度是BPTT算法的线性近似。它很容易实现，需要在PyTorch（Paszke等人，2019）中进行no\_grad操作或在TensorFlow（Abadi等人，2016）中进行stop\_gradient操作，并将时间和空间复杂度从BPTT的O（t）降低到O（1）。我们测试了在梯度中加入更多的项，但其性能比使用一步的要差。根据实验结果，单步梯度是可以接受的，可以通过MD进行反向传播梯度。



3 实验

在这一节中，我们提出了证明上述技术的实验结果。两个从全局信息和注意力机制中获益良多的视觉任务吸引了我们，包括语义分割（超过50篇使用注意力的论文）和GANs等深度生成模型（自SAGAN（Zhang等人，2019a）以来，大多数最先进的GANs都采用了自我注意力）。这两项任务都具有很强的竞争性，因此足以将Hamburger与自我注意进行比较。消减研究显示了Hamburger中MD的重要性，以及一步梯度的必要性。我们强调Hamburger在全局环境建模方面比自我注意在性能和计算成本方面都要优越。

3.1 消融实验

我们选择在PASCAL VOC数据集（Everingham等人，2010）上进行所有的消融实验，用于语义分割，并以best(mean)的形式报告在验证集上5次运行的mIoU。输出跨度为16的ResNet-50（He等人，2016）是所有消融实验的主干。我们采用3×3 conv与BN（Ioffe & Szegedy, 2015）和ReLU，将通道从2048个减少到512个，然后加入Hamburger，与语义分割中的流行关注点相同。关于详细的训练设置，请见附录E.1。  
**面包和火腿** 我们对汉堡的每个部分进行消融。移除MD（火腿）会导致性能最严重的衰减，证明了MD的重要性。即使只添加无参数的MD（只有火腿），性能也能明显改善。参数化也有助于汉堡包处理提取的特征。面包，尤其是上面包，对性能的贡献相当大。

**潜维d和r** 值得注意的是，d和r与mIoU测量的性能之间没有简单的线性关系，尽管d=8r是一个令人满意的选择。实验表明，即使是r=8也有很好的表现，揭示了它对全局环境的建模可以是非常便宜的。

**迭代K** 我们在评估阶段测试更多的优化步骤。一般来说，建议对训练和测试采用相同的K。对于CD和NMF来说，K=6已经足够了，而对于VQ来说，甚至K=1也是可以接受的。通常情况下，3∼6步就足够了，因为简单的MD的先验仍然是有偏差的，完全收敛会使其过拟合。这几个迭代很便宜，可以作为早期停止。

3.2 仔细观察HAMBURGER

为了理解Hamburger在网络中的行为，我们在PASCAL VOC验证集上可视化了Hamburger之前和之后的表示谱。输入和输出的张量被展开为RC×HW。图5显示了最大r奇异值的平方与展开后的矩阵总奇异值的累积比率。在经典的矩阵分解模型的结果中，由于低秩重建，通常会观察到一个截断的频谱。在网络中，Hamburger也促进了能量的集中，同时通过跳过连接保留了信息性的细节。此外，我们在图6中可视化了Hamburger前后的特征图。MD帮助Hamburger学习可解释的全局信息，方法是将无信息的通道清零，去除不规则的噪音，并根据上下文完成细节的处理

3.3 与注意力的比较

本节展示了基于MD的Hamburger在计算成本、内存消耗和推理时间方面相对于注意力相关的情境模块的优势。我们将Hamburger（Ham）与自我注意（SA）（Vaswani等人，2017）、DANet的双重注意（DA）模块（Fu等人，2019）、A2 Net的双重注意模块（Chen等人。2018b），APCNet的APC模块（He等人，2019b），DMNet的DM模块（He等人，2019a），CFNet的ACF模块（Zhang等人，2019b），报告参数和处理张量Z∈R1×512×128×128的成本见Tab. 3. 过多的内存使用是实际应用中与注意力合作的关键瓶颈。因此，我们也提供了NVIDIA TITAN Xp上的GPU负载和推理时间。总的来说，与注意力相关的全局上下文模块相比，Hamburger的计算量和内存都很轻。

3.4 精细分类

我们在PASCAL VOC数据集（Everingham等人，2010）和PASCAL Context数据集（Mottaghi等人，2014）上对Hamburger进行了基准测试，与最先进的关注点进行对比。我们使用ResNet-101（He等人，2016）作为我们的主干网。骨干网的输出跨度为8，分割头与消融实验相同。在消融研究中，NMF通常比CD和VQ更好（见表1）。因此，我们在进一步的实验中主要测试NMF。在下面的章节中，我们用HamNet来代表ResNet与Hamburger。在PASCAL VOC测试集和PASCAL Context验证集上的结果，见表4和表5。4和Tab. 5，分别说明。我们把所有基于注意力的模型都标上∗，其中不同的注意力组成了分割头。尽管语义分割是一项饱和的任务，而且大多数当代发表的作品都有近似的表现，但Hamburger显示出比以前最先进的注意力模块有相当大的改进。  
  
3.5 图像生成  
注意力在GANs等深度生成模型中呈现为全局上下文描述块。自SAGAN（Zhang等人，2019a）以来，大多数最先进的用于条件性图像生成的GAN都将自我注意力整合到其架构中，例如BigGAN（Brock等人，2018）、S3GAN（Luˇ ci ́ c等人，2019）和LOGAN（Wu等人，2019）。在ImageNet（Deng等人，2009）上具有挑战性的条件图像生成任务中，对基于MD的Hamburger进行基准测试是令人信服的。我们进行了实验，在ImageNet 128×128上比较Hamburger和self-attention。在特征分辨率为32×32的图像生成器和判别器中，用Hamburger和NMF ham代替了Selfattention，并命名为HamGAN-baby。与SAGAN相比，HamGAN在Fr ́echet Inception Distance（FID）（Heusel等人，2017）方面取得了明显的改进。此外，我们将Hamburger与最近开发的注意力变体Your Local GAN（YLG）（Daras等人，2020）进行比较，使用他们的代码库和相同的训练设置，命名为HamGAN-strong。在同一个TPUv3训练平台上，HamGAN-strong在FID方面有超过5%的改进，同时在总训练时间方面快了15%，在模块时间方面快了3.6倍（HamGAN为1.54次/秒，YLG为1.31次/秒，没有两个上下文模块为1.65次/秒，1000次迭代的平均值）。  
  
5 结论

本文研究了网络中长距离依赖关系的建模。我们将学习全局背景表述为一个低秩恢复问题。在这样的低秩表述的启发下，我们在研究透彻的矩阵分解模型的基础上开发了汉堡模块。通过对矩阵分解的目标函数的专业化，其优化算法所创建的计算图自然地定义了汉堡包的核心架构--火腿。Hamburger通过去噪和完善其输入学习可解释的全局背景，并重新调整频谱的浓度。令人吃惊的是，当谨慎地应对后向梯度时，即使是20年前提出的简单矩阵分解，在具有挑战性的视觉任务语义分割和图像生成中，也和自我注意一样强大，而且轻巧、快速、内存高效。我们计划通过整合位置信息和设计类似Transformer的解码器将Hamburger扩展到自然语言处理，为一步梯度技巧建立理论基础，或者找到更好的方法来区分MD，并在未来整合高级MD。