**解释双模和编码—解码器的Transformers的通用性自注意模型的可解释性**

**摘要**

Transformers越来越多地主导着多模态推理任务，比如视觉问题回答，是由于它们能够利用自注意力和共同注意力将信息情境化的能力，从而能够取得最先进的结果。这些注意力模块在其他计算机视觉任务中也同样发挥着作用，包括物体检测和图像分割。与其他只使用自我注意的Transformers不同，具有协同注意力的Transformers需要平行考虑多个注意力图，以突出与模型输入中的预测相关的信息。在这项工作中，我们提出了第一种方法来解释任何基于Transformers的架构的预测，包括双模式的Transformers和具有共同关注的Transformers。我们提供通用的解决方案，并将其应用于这些架构中最常用的三种：(i) 纯粹的自注意力机制，(ii) 自注意力机制与共同注意力机制相结合，以及(iii) 编码器-解码器注意力机制。通过演示，我们的方法优于所有由单模态可解释性改编的现有方法。我们的代码见：https://github.com/hila-chefer/ Transformer-MM-Explainability。

**1.绪论**

多模态Transformers可能会改变计算机的视觉实践方法。虽然最先进的计算机视觉模型通常被训练成特定任务的模型，推断出固定数量的标签，但Radford等人[28]已经证明，通过训练一个图像——文本模型，采用Transformers对每种模式进行编码，可以执行数十种下游任务，无需进一步训练（"零次"），其准确度与最先进的技术水平相当。随后，Ramesh等人[30]使用一个双模Transformers，以前所未有的性能生成符合未见领域中给定描述的图像。

这两个贡献出了以不同方式合并了文本和图像。第一种是用Transformer[40]对文本进行编码，用ResNet[15]或Transformer对图像进行编码，然后应用对称的对比性损失。第二种是将量化的图像表示与文本标记相连接，然后采用Transformers模型。也有许多其他结合文本和图像的方法[38, 21, 19, 18]。所有这些的共同点是，从两个输入到预测的映射包含两种模式之间的互动。这些互动往往对现有的针对基于注意力的模型的可解释性方法提出挑战，因为就我们所能确定的而言，所有现有的Transformers可解释性方法（例如[5，1]）都严重依赖于自注意力机制，而没有提供对任何其他形式的注意力机制的适应性，而这正是多模态Transformers所常用的。

另一类不限于自注意力的Transformer模型是Transformer 编码—解码器，即生成模型，其中模型通常从单一领域接收输入，并从不同领域产生输出。这些模型被用于一类新兴的物体检测[4, 49]和图像分割[43, 27, 42]方法中，也被广泛用于各种NLP（自然语言处理）任务，如机器翻译[40, 17]。例如，在这些物体检测方法中，特定位置和特定类别查询的嵌入与编码的图像信息交叉进行。

我们提出了第一个适用于所有Transformer架构的可解释性方法，并在三种最常用的Transformer架构上证明了其有效性：(i) 纯粹的自注意力机制，(ii) 自注意力机制与共同注意力机制相结合，以及(iii) 编码—解码器注意力机制。我们使用每个架构中的典范模型，并证明我们的方法比现有的Transformer可解释性方法更有优势，因为这些方法是由单一模式的起源改编的。我们的可解释性处方比现有的方法，如[5]，更容易实现，并且可以很容易地应用于任何基于注意力机制的架构。

**2.相关工作**

**计算机视觉中的可解释性**

解释计算机视觉算法通常需要合成一个热图，描述每个图像位置的计算相关度。这可以是依赖于类别的（对于每一个可能的标签），也可以是不依赖于类别的，在这种情况下，它只依赖于输入和模型。与以下大多数方法不同，我们的方法属于第一种类型。可解释性方法有多个系列，包括基于显著性的方法[8, 34, 23, 48, 44, 47]，考虑激活的方法[10]，使用前向传递或后向传递[45]，基于扰动的方法[11, 12]，以及基于Shapley-values（夏普利值）的方法[22, 6]。后者享有明确的理论动机。通过深度泰勒分解理论[24]，基于归属的方法也得到了理论上的支持。这种方法从顶层向后递归地分配相关度，使相关度之和保持固定。LRP方法[3]，就是这样一个突出的方法。由于LRP和大多数变体[25, 33, 22]是不分等级的[16]，因此引入了特定等级的扩展[13, 16, 14]。

基于梯度的方法直接考虑损失相对于每层输入的梯度，通过反向传播来计算。这方面的例子包括类别无关的方法[33, 37, 35, 36]。一个相关的针对类的方法是Grad-CAM方法[32]，该方法考虑了顶层的输入特征与类相关的梯度。

**Transformers的可解释性**

大多数解释Transformers的尝试都直接采用了注意力图。然而，这忽略了中间的注意力分数，以及Transformers的其他组成部分。正如Chefer等人[5]所指出的，每个注意力机制头中的计算混合了查询、键和值，仅考虑查询和键的内积是不能完全体现的，这就是所谓的注意力机制。

Voita等人[41]应用LRP来捕捉每个Transformer块中注意力机制头的相对重要性。然而，这种方法并没有将相关性分数传播回输入，以产生一个热图。

Abnar等人[1]提出了一种将多个层次的注意分数结合起来的方法。他们提出了两种方法：注意力展开和注意力流动。第一种方法是沿着成对的注意力图中的备选路径线性地结合注意力。在[5]中显示，这种方法不能区分对决策的积极和消极贡献，导致在应该取消相关分数的情况下，相关分数在各层中的累积。注意力流方法被表述为同一个成对注意力图上的最大流问题。虽然在[1]中显示，在特定的场景下，它在一定程度上优于rollout，但这种方法太慢，无法支持大规模的评估。

与这些方法相比，Chefer等人[5]对Transformer模型的所有组件内的信息传播进行了全面的处理，该模型通过所有的层将信息从决策层反向传播到输入层。该解决方案是基于分层相关性传播[3]，对自注意力层进行梯度整合，并被证明对单一模式的Transformer编码器非常有效，如[9].然而，该方法并没有为自注意力以外的注意力模块提供解决方案，因此不能为所有Transformer架构提供解释。

**计算机视觉中的Transformers**

Transformer技术在双模任务中已变得越来越普遍，如图像字幕和基于文本的图像检索。我们区分了依赖自注意力的网络，如VisualBERT[18]和Oscar[19]，以及那些也采用共同注意力模块的网络，如LXMERT[38]和ViLBERT[21]。我们的方法为这两种类型提供了合适的可视化。

我们的方法也提供了第一个完整的解决方案，就我们所能确定的而言，Transformer编码—解码器[40, 29, 17]，在计算机视觉中越来越普遍。在DETR基于Transformer的检测方法[4]中，图像由变换器编码器编码，获得的信息与基于位置和类别的查询一起被共同关注。我们的方法也可以应用于基于编码器的视觉Transformer，如用于图像识别的Transformer[7, 9, 39]，以及用CNN解码器进行图像分割[46]。然而在这种情况下，现有的Transformer可解释性方法也可以应用。

**3.办法**

我们的方法是利用模型的注意力层为网络中的每个输入模态之间的互动产生相关性图。在这项工作中，我们专注于图像和文本的互动，以及生成模型的注意模块，即编码—解码器注意。然而，我们的方法很容易适用于任何基于Transformer的架构，也可以被推广到处理两种以上的模式。在下文中，我们将讨论该方法在两种模态假设下的传播规则，例如为简单起见，文字和图像，然后详细描述如何将我们的方法应用于本工作中使用的每种模型类型。

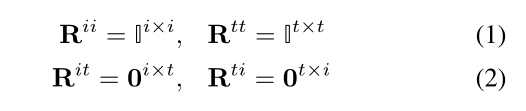
*t*，*i*分别为文本和图像输入标记的数量。为了简化符号，我们使用相同的符号（*t*，*i*）来标识与这两个领域相关的变量。多模态注意网络包含四种类型的输入标记之间的相互作用：A*tt*和A*ii*分别是文本和图像标记的自注意力机制的相互作用。A*ti*和A*it*是多模态注意力机制的相互作用，其中A*ti*代表图像标记对每个文本标记的影响，A*it*代表文本标记对每个图像标记的影响。

根据所描述的注意力机制交互作用，我们为每个交互作用构建一个相关图，即R*tt*，R*ii*用于自注意力机制，R*ti*，R*it*用于双模注意力机制。

该方法通过对注意力层的前向传递来计算相关性图，每一层都使用我们将在下面的小节中描述的更新规则对聚合的相关性矩阵所提供的作用。

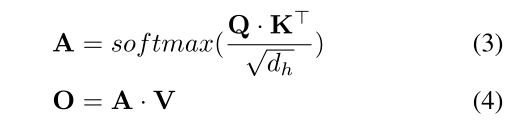
**相关性初始化**

在注意力机制操作之前，每个标记都是自成一体的。因此，自注意力机制的互动是以身份矩阵初始化的。对于双模互动，在注意力层之前，每个模态都是独立的，不包含来自其他模态的背景，因此相关性图被初始化为零。



**相关性更新规则**

随着注意力层对标记的语境化，我们的方法修改了受标记嵌入的混合物影响的相关性图。回顾一下[40]中提出的注意机制。

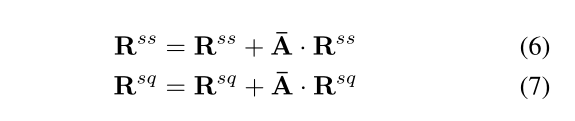


其中（）表示矩阵乘法，O是注意力模块的输出，Q是查询矩阵，K，V是关键字和值的矩阵。*h*是头的数量，*dh*是嵌入维度，表示领域和每个领域的标记数量，即注意力机制发生在*s*查询标记和*q*关键标记之间。请注意，从图1中可以看出，对于自注意力层来说，*s*=*q*，Q、K、V都是输入到注意力单元的映射，而在共注意力机制中，Q是输入的映射，K、V是来自其他模态的背景输入的映射。A是注意力图，它直观地定义了来自*s*，*q*的每一对标记之间的连接。由于注意模块后面有一个残余连接，如图1所示，我们通过将每一层的作用加到汇总的相关度中来累积相关度，类似于[1]中加入身份矩阵来考虑残余连接。

我们的方法使用每个注意层的注意图A来更新相关性图。由于每个这样的地图是由h个头组成的，我们遵循[5]，使用梯度来平均各个头。请注意，Voita等人[41]的研究表明，注意力头的重要性和相关性不同，因此在各头之间的简单平均会导致扭曲的相关性图。然后，我们方法的最终注意力图定义如下：



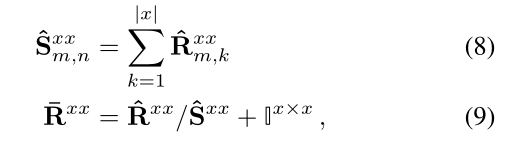
其中⊙是哈达玛德积，中的yt我们希望可视化的t类的模型的输出，Eh是整个头维的平均值。按照[5]，我们在平均化之前去除负作用。对于满足的自注意力层，受影响的聚合相关度分数的更新规则是：



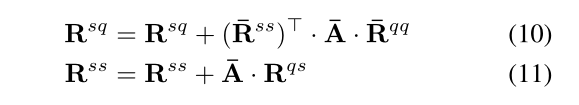
在公式6中，我们通过应用矩阵乘以聚合的自注意力矩阵Rss，解释了标记在以前的注意力层中已经被置于语境中考虑，正如在[1，5]中所做的那样。对于公式 7，注意前面的双模注意力层将来自*q*的语境插入到*s*中，因此，当自我注意混合来自*s*的标记时，它也混合来自*s*的每个标记中的语境*q*。R*sq*体现了前几层背景的混合。因此，我们从自注意力机制的过程中计算增加的语境。

对于，其中应用了双模注意力机制，关联累加器的更新规则包括对自注意力矩阵R*xx*, *x*∈{*s*, *q*}的归一化。由于我们初始化了R*xx* ，而公式 6累积了每一层的关联矩阵，所以我们可以将聚合的自注意力矩阵R*xx*看作是由两部分组成的矩阵，第一部分是初始化时的单位矩阵，而第二部分是由各层自注意力的聚集所形成的矩阵。

由于公式5使用梯度来平均各头，所以的值通常会减少。我们希望既考虑到每个标记对自身的影响，又考虑到自注意力机制的语境化。因此，我们将中的每一行归一化，使其总和为1。直观地说，中的第*i*行披露了每个标记对第*i*个令牌的自注意力值，而单位矩阵将每个标记对自身的该值设为1。 因此：



其中/代表矩阵逐个元素的除法。在上面，我们将中的每一行归一化，即用该行的每个元素除以该行的和。接下来，我们为双模注意力单元定义以下聚合规则：



公式10说明了这样一个事实，即每种模式的标记在以前的注意力层中已经被语境化了，方法是应用矩阵乘以归一化的聚合自注意力矩阵, 。

对于公式11，注意到前面的双模注意力层整合了两种模态的嵌入，因此，当用*q*将*s*联系起来时，*q*也包含来自*s*的信息，体现在Rqs中。

注意，上述规则是根据模式*s*∈{*i*, *t*}的输入和模式*q*∈{*i*, *t*}的语境来描述的，也就是说，这些规则是对称地应用于图像和文本这两种模式的。

**3.1. 获得分类相关度**

为了进行最终的分类，基于Transformer的模型通常会考虑[CLS]标记，这是一个被添加到输入标记中的标记，构建了所有输入标记的一般表示。为了检索分类任务中每个标记的相关度，可以考虑相应相关度图中[CLS]标记对应的行。例如，假设[CLS]标记是文本模式中的第一个标记，要提取每个文本标记的相关度，应考虑R*tt*的第一行，要提取图像标记的相关度，应考虑R*ti*的第一行，它描述了[CLS]标记和每个图像标记之间的联系。

**3.2. 对注意类型的适应性**

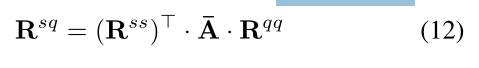
在这项工作中，我们在基于Transformer的网络中使用的三种不同类型的注意机制上检验我们的方法。架构和匹配的传播规则在图2中得到了可视化。第一种架构类型是多模态的Transformer，其中两种模态被串联起来，并由[SEP]标记分隔[18, 19]，如图2（a）所示。这样的网络只使用自我注意来确定模态的语境，即只使用公式6。由于该模型是基于纯粹的自注意力，我们产生了一个相关图R(*t+i*,*t+i*)，它定义了各模态之间以及各模态内部的联系。为了可视化与分类相关的标记，我们应该考虑R(*t+i*,*t+i*)中与用于分类的标记相对应的那一行。这一行Rcls(*t*+*i*)产生了每个图像标记和每个文本标记的相关度分数。

第二种类型是多模态注意力网络，它包括共注意力模块，将每种模态与其他模态联系起来[38, 21]，如图2（b）所示。这样的网络需要上述所有的传播规则，针对每一种模式。为了产生分类的相关性，我们只需遵循第3.1节的例子，因为如图2(b)所示，[CLS]标记在这种情况下是文本模态的第一个标记。

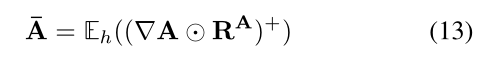
第三种也是最后一种类型是生成模型，其中有一个输入模式，而输出来自不同的领域[4, 49, 43, 27, 42, 40, 17]，这在图2(c)中得到了可视化。这种网络包含一个利用输入的自注意力的编码器和一个解码器。解码器有两种类型的输入，第一类是保持不变的编码数据，第二类是来自解码器领域的输入。解码器在解码器领域的标记上利用自注意力，然后由一个共注意力层将它们与编码器的输出联系起来。为了说明问题，在这种情况下，相关性更新规则如下：用*e*表示编码器的标记，用*d*表示解码器的标记。相关性矩阵为：R*ee*，R*dd*为自注意力的互动，R*de*为解码器的标记和编码器的标记之间的双模式互动。请注意，由于编码器没有被语境化，我们没有相关性矩阵R*ed*。编码器对R*ee*的自注意力计算只需遵循公式6。对于解码器的自注意力计算，我们应用公式6，7。对于解码器的双模注意力计算，我们按照公式10来计算编码器和解码器的自注意力。请注意，公式11是不相关的，因为我们没有R*qs*=R*ed*的相关度图。为了在这种情况下提取相关性，我们考虑相关性图R*de*。在这项工作中，我们使用一个物体检测模型作为我们的典范编码—解码器架构。对于这样的模型，来自*d*的每个标记都是一个查询，代表输入图像中的一个物体。为了产生每个图像区域与被检测物体*j*的相关性，我们应该考虑R*de*的第*j*行，它对应于第*j*个检测。包含每个编码器标记的相关度分数，在这种情况下是一个图像区域。

**4. 基准线**

我们专注于那些在可解释性文献中常见的方法，并且适用于我们在这项工作中报告的广泛测试。按照[5]，我们提出了三类基线：注意力图基线、梯度基线和相关性图基线。我们的注意图基线是原始注意力和展开。原始注意力只把最后一层的注意力图视为相关图，例如，R*tt* = A*tt*，其中A*tt*是最后一个文本自注意力图。第二种是rollout算法，对所有的自我注意层都遵循[1]。由于rollout基线完全基于自注意力，为了区别于原始注意力，我们对R*sq*，*s*，*q*∈{*t*，*i*}采用如下方法：



其中，R*ss*、R*qq*是通过rollout计算的自注意力相关性，是最后的双注意力图。对于我们的梯度基线，我们使用[5]中描述的Grad-CAM[32]适应，即我们检查最后一个注意力层，并对注意力图的头部进行Grad-CAM。最后，我们的相关性图基线包括部分LRP，遵循[41]，它使用最后一个注意层的LRP相关性值来平均整个头，以及[5]中描述的Transformer归属方法。[5]中的方法对所有注意层采用公式5，以便在各头之间进行平均，具体方法如下：



其中与公式5相比，唯一的区别是[5]使用了A的LRP[3]相关性值，即RA，而不是像公式5那样使用原始注意力图。此外，[5]对所有自注意力层使用公式6。对于非自注意力层，我们的版本[5]采用最后一张注意力图，并使用公式13对各头进行平均。请注意，应用我们的方法只需要对注意力模块设置一些简单的联系，而LRP则需要对所有网络层进行自定义实现。

**5. 实验**

我们的实验包括三个基于Transformer的模型，每个模型代表我们在本工作中提到的三种类型的架构之一。每个架构的图示见图2。此外，为了与以前的工作[5，1]进行比较，我们还考虑了ViT[9]，这些方法是在相同的环境下构思的。每个模型的相关度传播遵循第3.2节。

我们研究的第一个模型是VisualBERT[18]，它代表了一个基于自注意力机制的架构，第二个模型是LXMERT[38]，它代表了一个结合自注意力机制和共注意力机制的架构，在一个Transformer编码器中用于两种模式。

对于这两个模型，我们分别对每个模态进行正、负扰动测试，以评估方法产生的相关矩阵的质量。我们使用视觉问题回答[2]任务来测试解释，因为这个任务要求模型表现出对两种输入模式和它们之间联系的理解。

扰动测试如下：首先，使用预先训练好的网络从VQA数据集的验证集中随机抽取10,000个样本来提取相关性图。其次，我们逐渐移除特定模式的标记，并测量网络的平均top-1准确性。在正向扰动中，标记被从最高相关性到最低相关性移除，而在负向扰动中，标记被从最低到最高移除。在正扰动中，我们期望看到性能的急剧下降，因为这表明被移除的标记对分类得分很重要。在负扰动中，一个好的解释会保持模型的准确性，同时删除与分类无关的标记。在这两种情况下，我们测量曲线下面积（AUC），以评估模型准确性的下降。

我们注意到，在所有的扰动测试中，即使去除每种模式的100%的标记，准确率也没有达到0%。这是由于来自其他模态的输入保持不变，因此模型可以依靠单一模态来提供合理的答案。

注意LXMERT[38]的图像扰动测试结果，如图4(a,b)所示，与其他方法相比，我们的方法有明显的优势。对于负扰动，使用我们的方法的AUC是最大的，有相当大的优势，即使在去除80%以上的图像标记后，准确率也能得到很好的保留，而对于正扰动，注意准确性非常急剧下降，AUC也很低。

从图2(b)可以看出，LXMERT[38]的[CLS]标记是文本模态的第一个标记，因此根据第3.2节，Rti是在图像扰动情况下用于提取相关度的地图。由于Rti是一个多模态的相关图，图像扰动测试最好地证明了使用我们的方法比所有现有的方法的优势，这些方法在评估共注意力模块的相关度方面有缺陷。

对于图4(c,d)中描述的LXMERT[38]文本扰动测试，注意到根据第3.2节，我们可以看到R*tt*是一个自注意力图，其中主导的更新规则是公式6。这个规则与Transformer attribution[5]基线采用的规则相同，除了公式13中的头部平均化。因此，我们提出的方法和[5]中描述的方法之间的主要区别是在头部平均化过程中选择使用LRP[3]。这使得两种方法的结果非常相似。为了完整起见，我们在补充文件中提供了我们的方法在加入LRP时的结果，正如公式13所做的。其余的方法都远远落在后面。

图4展示了我们的方法和Transformer attribution[5]的典型结果。其余的方法没有竞争力，它们的匹配样本在补充文件中列出。可以看出，文本的结果是相似的，正如定量结果所预测的那样。我们的图像注意力结果比基线方法的结果更集中于相关的图像部分。

请注意，由于VisualBERT[18]是基于纯粹的自注意力，我们的方法和Transformer attribution[5]方法之间的差异源于是否使用LRP[3]进行公式5中的头部平均化的选择，类似于LXMERT[38]的文本（但不是图像）扰动测试。从图5中可以看出，我们的方法优于所有的方法，并取得了与[5]非常相似的结果，在某些情况下，如文本扰动测试，甚至比[5]的结果要好很多。这表明，即使对于纯粹的自注意力架构，也没有必要使用LRP[3]。

我们实验的第三个模型是DETR[4]，它是一个编码—解码器模型，如图2（c）所示。我们使用预先训练好的DETR模型和ImageNet的预训练骨干ResNet-50，它是为MSCOCO[20]数据集上的物体检测而训练的。重要的是，这个模型只被训练用于物体检测，即为输入图像中的每个物体产生边界框和分类。为了评估不同的可解释性方法，我们的测试在MSCOCO[20]验证集的5000个样本上使用每个方法来产生分割掩码，即我们认为每个方法的输出都是一个分割掩码。我们首先对查询进行过滤，只包括分类概率高于50%的查询，然后采用大津的阈值处理方法[26]来分离分割的前景和背景。全部细节见补充说明。

我们生成的分割掩码将DETR预测的界线盒可视化，因此应该注意到，生成的掩码本质上取决于相应界线盒的质量，也就是说，当预测的界线盒不充分时，自然，为其生成的掩码至少同样不准确。此外，由于可解释性方法不是以产生分割图为目的，它们通常不会输出连续的掩模，而且大津阈值也可能在产生的掩模中产生 "洞"。由于上述原因，我们将用于MSCOCO评估的最小IoU从0.5降至0.2，这对所有的方法都大有裨益，我们将MSCOCO分割评估的结果呈现在所产生的边界框足以生成分割掩码的类别上，例如，我们不呈现小物体的结果1。从表1中可以看出，我们的方法在很大程度上超过了其他方法。1，我们的方法以非常大的优势胜过其他所有的方法，这表明我们的新配方对于非自觉性架构是必要的。请注意表1中的关联性。1中，DETR的边界框评估和我们的分割方法之间的相关性。掩码的可视化见图6。

最后，为了将我们的方法与现有的单模态基线进行比较，我们提出了对ViT-Base[9]的正、负扰动测试，如[5]所做的那样。如前所述，由于ViT-Base[9]是一个单模态的Transformer编码器，我们的方法与[5]的Transformer归属方法的唯一区别是在公式5中使用了LRP[3]，如公式13所示。因此，从Tab. 2，我们的方法和[5]中提出的方法之间的差异是非常缓和的，这也说明LRP[3]可以被删除。Tab. 2还显示了在公式5中使用目标类而不是预测类进行梯度传播时的性能改进，正如[5]中所说的，这表明我们的方法能够产生特定类别的可视化。

**消融研究**

我们在补充文件中提出了我们方法的三种变化，证明了我们的归一化（公式8,9）的有效性，在我们所有的规则6、7、10、11中聚合的必要性，以及对双模式规则10进行自我注意更新的必要性。

**6. 结论**

Transformers在计算机视觉中发挥着越来越大的主导作用，图像—文本的Transformers和执行任务的Transformers，其输出域比分类器提供的标签更复杂，呈现出突破性的成果。为了调试这类模型，以及支持下游任务，以及对模型可解释性的要求越来越高，需要有完整和准确的可解释方法。然而，目前关于Transformers的可解释性文献是有限的，过度关注于纯注意力图，而缺乏处理共注意力图的方法。

我们的方法仔细跟踪了注意力地图的演变和混合。它提供了一个通用的处方，适用于我们所知道的所有注意力模型。从经验上看，它在不同的Transformer架构和评估指标上都优于现有的方法。在某些情况下，当自注意力很突出时，Chefer等人[5]的最新方法是唯一能提供类似结果的方法。然而，在大多数实验中，我们的方法以非常可观的幅度领先于所有方法。