UniT: Multimodal Multitask Learning with a Uniﬁed Transformer使用统一Transformer进行多模式多任务学习

代码: [https://mmf.sh](https://link.zhihu.com/?target=https%3A//mmf.sh)

**论文动机：是将transformer应用于跨领域的不同任务。**

在这项工作中，建立了一个统一模型UniT，它将图像和/或文本作为输入，并在从视觉感知和自然语言理解到视觉和语言联合推理等多个任务上联合训练。UniT由transformer编码器和transformer解码器组成，前者将每个输入模态编码为一连串的隐藏状态（特征向量），后者将编码后的输入模态进行解码，然后将特定任务的输出头应用于解码器的隐藏状态，对每个任务做出最终预测。

贡献如下：

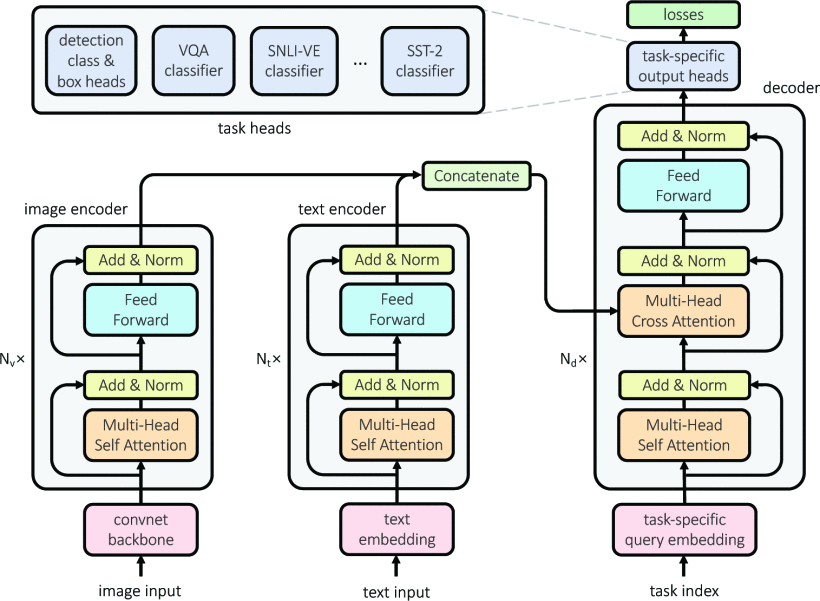
1.提出UniT，一个统一的transformer编码器解码器架构，在一个参数较少的单一模型中处理多个任务和领域。

2.联合学习了视觉和文本领域中最突出的任务及其交叉点，即目标检测、视觉问答（VQA）、视觉蕴含和GLUE基准测试中的自然语言理解任务，包括QNLI、MNLI、QQP和SST-2。作者表明，这些不同的任务可以同时学习，并在他们的训练方案下正常收敛。

3.通过对各种任务的分析，表明多模态任务，如VQA（视觉回答）和视觉衔接，都受益于单模态任务的多任务训练。

UniT模型的概述它通过一个统一的transformer编码器-解码器架构联合处理不同领域的各种任务。

1. 输入:该框架可以接受两种或更多的输入模态,比如图像I和语音S。每个模态先通过自己的embedder（嵌入器）,得到对应模态的嵌入向量,比如图像嵌入aI和语音嵌入bS。后续分别介绍了文本输入与图片输入的流程。
2. Encoder（编码器）:它使用统一的Transformer encoder对两个模态的嵌入向量进行编码。首先对aI和bS拼接得到[aI; bS],然后输入Transformer encoder。编码后得到融合的向量c,它包含两种模态的信息。



1. Decoder（解码器）:它也使用统一的Transformer decoder。根据任务不同,它生成不同的输出,比如语言描述yL或图像标注yI。
2. 损失函数: 它基于任务的不同使用不同的损失函数,比如交叉熵损失用于语言生成,平滑L1损失用于标注回归等。对语言生成任务使用交叉熵损失。
3. 处理与该任务相关的输入和输出：来自transformer解码器的表示特征被传递给一个特定任务预测头，比如一个简单的两层分类器，以输出最终的预测。

即UniT由transformer编码器和transformer解码器组成，transformer编码器使用图像编码器对视觉输入进行编码，使用文本编码器对语言输入进行编码，将每个输入模态编码为一系列隐藏状态（特征向量），并使用共享的transformer解码器（联合解码器）在编码后的输入模态上对每个任务的查询进行嵌入，然后在解码器的隐藏状态使用特定任务预测头，对每个任务进行初始预测，以生成每个任务的最终输出。整个模型经过端到端的联合训练。每个任务都有损失。