**多模态Transformer**

1. **介绍**

人类语言不仅拥有口语，还拥有视觉(面部特征)和声学(语调)形式的非语言行为，因此人类语言是多模态的。这些丰富的信息为我们理解人类的行为和意图提供了好处。然而，不同模态之间的异质性往往增加了分析人类语言的难度。例如，音频和视频流的接收器可能会随着接收频率的变化而变化，因此可能无法在它们之间获得最佳映射。例如，皱眉可能与过去说过的悲观的话有关。也就是说，多模态语言序列通常表现出“不对齐”的性质，并且需要推断不同模态之间的长期依赖关系。

Transformer中的self-attention可以很好的学习到输入序列的依赖关系，但是人类多模态语言时间序列既不像单词嵌入那样具有良好的代表性，也不像单词嵌入那样离散，每种模态的序列具有非常不同的频率。因此，像NMT Transformer那样使用编解码器结构，显式的将一种序列转换成另一种序列是非常困难的。所以，多模态Transformer不使用编解码结构，只通过注意力学习不同模态时序序列的潜在关系。

1. **模型结构**

按照输入到输出的顺序该模型主要由时间卷积、位置嵌入、跨模态Transformer、自注意力Transformer和结果预测组成，如图1所示。

图示

描述已自动生成

图1 模型总体结构

1. **时间卷积**

一维时间卷积的作用是为了确保输入序列的每个元素对其相邻元素有足够的感知：



其中为模态的序列长度，为三种模态的输入特征序列，是模态的卷积核大小。由于每个模态的维度是不同的，会导致在后续的跨模态注意块中无法进行运算。因此需要时间卷积将不同模态的特征投影到相同的维度d，这样就可以在后续的跨模态注意模块中执行点积运算。

1. **位置嵌入**

与原始的Transformer一样，为了使序列能够携带时间信息，将位置嵌入添加到卷积层的输出上，使得输入到跨模态Transformer的信息是有时间顺序的。



1. **跨模态Transformer**

图示

描述已自动生成跨模态Transformer使一种模态能够接受来自另一种模态的信息。下面演示将视觉(V)信息传递给语言(L)信息的例子，用V → L表示，由于前面输入信息经过了时间卷积的处理，因此每个模态的所有的维度都是固定d。跨模态Transformer的作用可以理解为使用源模态V对目标模态L的信息进行重构，以学习到跨模态关联的有意义的信息。每个跨模态Transformers由D层跨模态注意块组成，如图2所示。

图2 跨模态Transformer结构

形式上，跨模态transformer按照层进行前馈计算的公式为：







其中位于第i层的的多头版本。为​计算不同模态之间的跨模态注意。融合跨模态信息的一个好的办法是提供一种跨模态的潜在适应，即从β到α。其中α代表视觉信息，β代表语言信息。在第i层跨模态注意块中，跨模态注意计算的输入为第i-1层跨模态注意块的输出然后再经过层归一化得到的，即，输入为源模态经过层归一化得到，即。从β到α的潜在适应表示为跨模态注意：



图示

描述已自动生成其中查询(query)来自模态α，键(key)和值(value) 来自模态β。具体的说，上式中按比例缩放的，其第(i,j)项由模态α的第i个时间步长对模态β的第j个时间步长的关注度。因此的第i个时间步长是的加权总和，其权重由中的第i行确定。式(2.6)为单头跨模态注意力，其计算过程如图3所示。

图3 不同模态序列的跨模态注意力计算过程

在计算完跨模态注意之后，进入位置前馈层。并且基于先前关于transformer的工作，在跨模态注意块中添加了残差连接以及在跨模态注意计算的前后添加层归一化操作，以组成完整的跨模态注意块。每个模态通过多头跨模态注意模块，与其他模态的底层信息进行交互不断更新其序列。在跨模态注意的级别上看，来自源模态(β)的低级信号被变换成一组不同的键和值与目标模态(α)进行交互。跨模态transformer能够学习到跨模态关联的有意义的信息。

1. **自注意力Transformer及预测**

作为最后一步，将来自共享相同目标模态的跨模态transformers的输出连接起来，以产生。例如，。它们中的每一个都通过自注意力transformer模型来收集时间信息以进行预测。最后，提取自注意力 transformer模型输出的最后一个元素并将其连接起来，然后通过全连接层进行预测。

1. **实验**

表格

描述已自动生成实验使用的数据集分别为CMU-MOSI、CMU-MOSEI、IEMOCAP。并分别将每种数据集划分为对齐与未对齐的。由于数据集较大，只展示划分好的数据集截图。

图4 划分数据集截图

文本

描述已自动生成文本

描述已自动生成实验结果分别如下：

图5 CMU-MOSI实验结果。左图为对齐的，右图为未对齐的。

文本

描述已自动生成文本

描述已自动生成图文本

描述已自动生成文本

描述已自动生成6 CMU-MOSEI实验结果。左图为对齐的，右图为未对齐的。

图7 IEMOCAP实验结果。左图为对齐的，右图为未对齐的

注：本次实验参考的是Multimodal Transformer for Unaligned Multimodal Language Sequences(Tsai et al. 2019 ACL)