TCAN论文阅读笔记

1. 引言部分

作者的目的是寻找一种架构，既能替代递归网络，又能吸收前馈模型的优点。

针对这个目的提出时间卷积注意力网络TCAN，网络包含两部分内容：**时间注意力TA**,用于捕捉序列内部的相关特征。**增强残差ER**，提取浅层重要信息并传递到深层。

1. 简介部分

三种常用的模型 for sequence modeling

Convolutional network

Recurrent network

Attention mechanisms

The problem of sequence modeling

作为任务的输入，序列在时间步 t 的数据点是基于之前的数据点的。

任意两个数据点之间可能存在相关性。

那么对于模型就至少有两个要求，一个是能够**表征因果关系**，一个是能够学习数据点之间的**条件相关性**。

在此之前，已经有人对递归模型和前馈模型做了一些尝试。

**多层感知机MLP**：为每个输入的特征独立训练参数，参数无法共享。

**时间延迟神经网络TDNN**：在每个时间步应用相同的卷积核实现了参数共享，但是记忆范围有限，无法处理数据密集的任务。

**RNN**：使用相同的参数处理从开始到当前时间步的输入，每个输出成员都是前一个输出成员的函数，且输入长度可变，满足**因果关系**和**条件相关性**。

Two shortcomings

后续时间步必须等待前面的完成，无法实现训**练评估并行化**。

更多关注附近信息，**忽略长距离信息**。

针对上述问题，近些年也出现了一些模型

TCN：无法学习序列中远距离位置的依赖关系。无法提取输入的内部相关信息。

TrellisNet：跨深度权重绑定，将输入直接注入深层，解决上述问题。

注：上述两个均为同一实验团队分别在2018年和2019年提出的，后者确实提高了性能，但是整个模型本身相较于TCN又大又慢。当然还有transformer，GPT-2等模型，性能也很好，但是同样规模太大。

综上，作者提出TCAN

一是利用膨胀因果网络模拟**因果关系**。

二是采用自注意力机制来提取内部相关信息并学习**远距离位置**依赖关系。

1. 模型结构

Sequence Modeling

Supposed an input sequence：

X\_{1:T} = x1,x2......xT length = T

Our target sequence:

Y\_{1:T} = y1,y2......yT length = T

We need find a function satisfies：

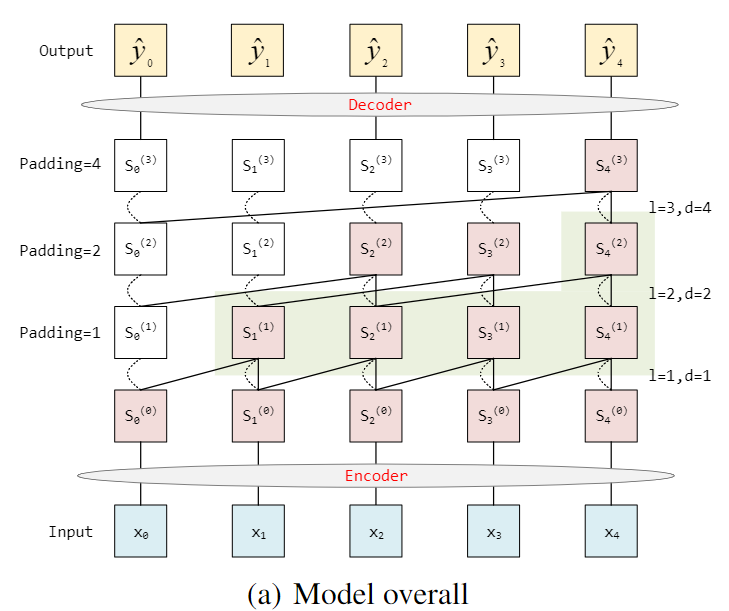
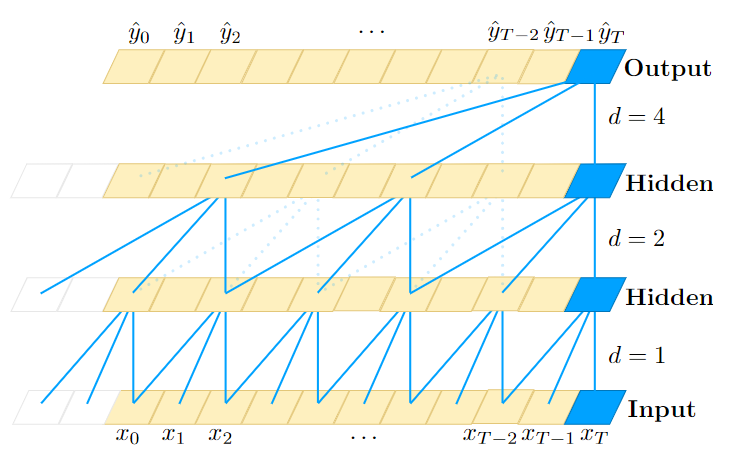
Y\_{1:T} = SeqMod(X\_{1:T})

Two constraint：

Y{:T}应该满足因果，它是X{:T}的函数，应防止未来信息X\_{t+1:T}

输入输出序列长度相同

A similar structure as TCN which contains encoder and decoder



**At the beginning** of the model：

X\_{1:T} = x1,x2......xT

to



T：the length of the sequence

0：0-th layer(the first hidden layer’s input)

将输入的符号序列（例如，单词或字符）转换为连续的向量表示。这些向量表示能够捕捉到输入序列中每个元素的语义和结构信息。

**Then** apply the different kernel sizes

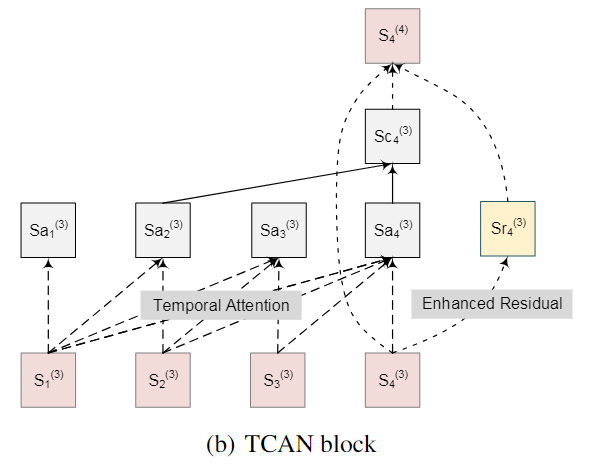
在L层隐藏层中应用不同卷积核大小的膨胀因果卷积。

**After the final** hidden layer

The decoder generates an output sequence



在第 t 个时间步和 L+1 层的中间变量通过四个步骤计算：



1、 passed through Temporal Attention



表示包含时间步 t 之前信息的中间变量

2、apply causal convolution on it



表示因果卷积的输出。因果块可以堆叠多层。

为了保持每层相同的长度，

在左侧添加长度为的零填充，

通过这种方式将左侧相关信息逐渐累加到右侧。

3、before feature maps being passed through the activate function

在通过激活函数获取之前，将三个成分相加：



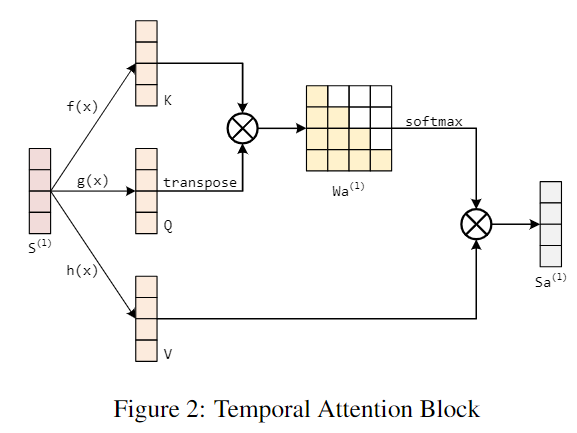
其中表示增强残差



4、通过在深度和时间上堆叠 L 层的TCAN块来构建一个完整的TCAN网络

为了使网络有足够的感受野，设置膨胀的大小随着网络的深度指数级增加：例如第 L 层，膨胀大小为 d = 2^L

**Temporal Attention**



TA层可以描述为将前几个时间步的影响整合到当前时间步。

使用三个不同的线性变换将映射到三个不同的向量：

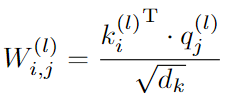
Keys：

Queries：

Value：

Dimension：

For getting the weigh matrix：

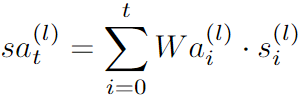


To achieve the purpose of not using **further information**：

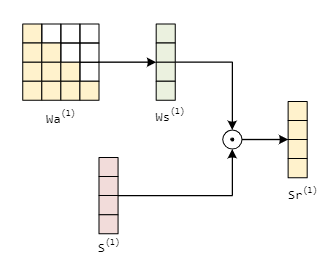


Then apply a **softmax function** to normalization

Final get the **weighted output**：



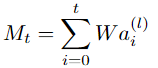
Enhanced Residual



增强残差利用了TA模块得到的权重矩阵

对权重矩阵的每一行的权重进行求和，来反应每个时间步的重要性程度

那么可以得到另一个权重向量：



Then after the Hadamard product we can get the 

注：

PPL（perplexity）：困惑度

自然语言处理中一个常用的评估语言模型的指标，反应一个语言模型对给定文本序列的预测能力。



N是测试数据集中的总词数，是模型给定前n个词的情况下预测第i个词的概率。

困惑度表示模型在平均情况下预测一个词所需的困惑程度。通常情况下，困惑度越低，表示模型的性能越好。