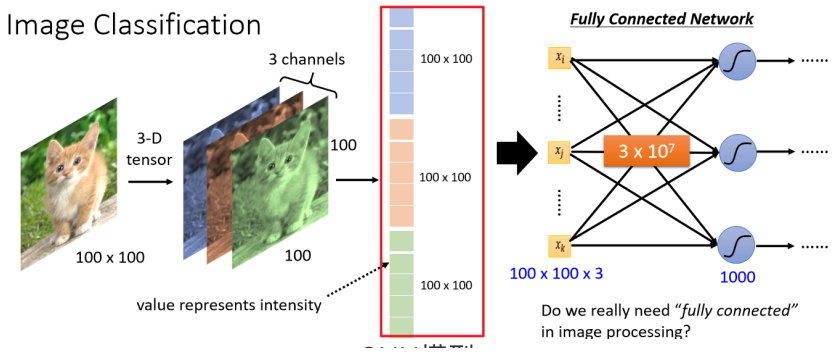
Transformer相关——（2）Seq2Seq模型

发表于 2021-08-16 | 更新于 2021-08-17 | 深度学习
| 字数总计: 862 | 阅读时长: 3分钟 | 阅读量: 1991

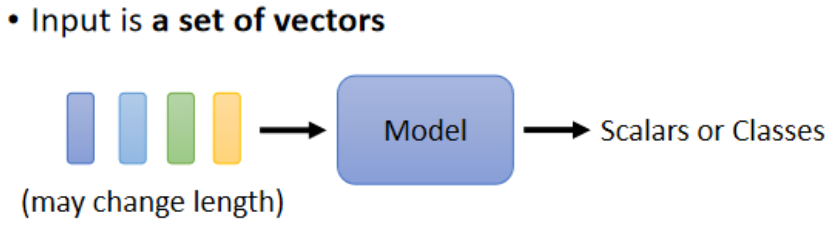
Transformer相关——（2）Seq2Seq模型

引言

上一节介绍了Encoder-Decoder框架，基于该框架设计的模型输入可以是**一个向量**（如最原始CNN的输入是图像经过flatten后的向量，下图红框部分，往往将数据处理成固定大小的向量作为输入）：

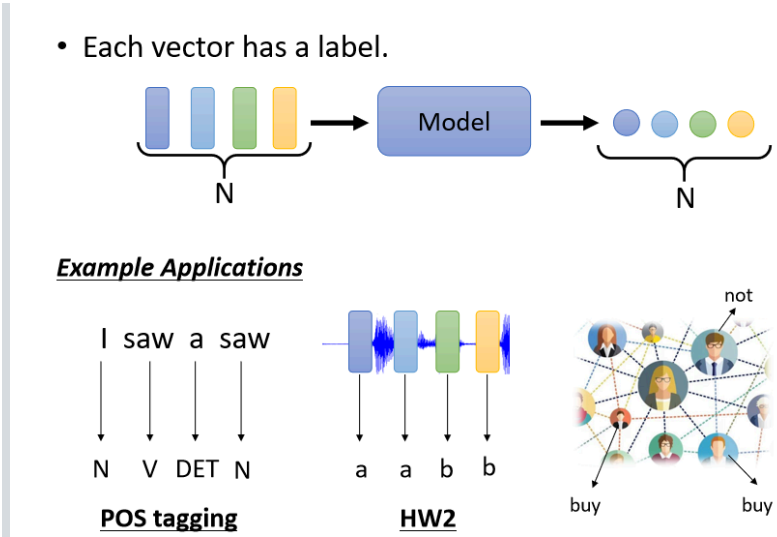


当面对更复杂的问题时，比如说面对一些具有时序特征的向量序列、具有顺序特征的序列，其表示成序列后，长度事先并不知道，那么为了适应这种输入是**多个向量**，而且这个**输入向量的数目是会改变的**的场景，需要设计新的模型。

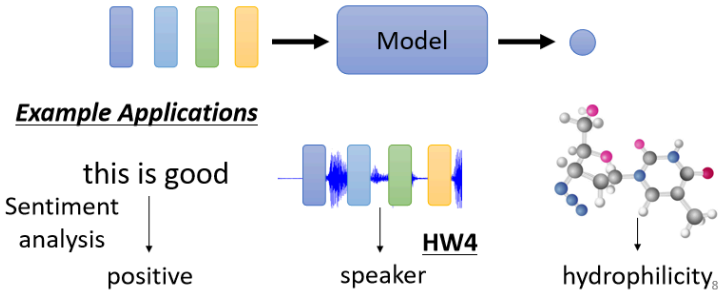


当输入是**多个向量**时，Decoder的输出可以有以下三种形式：

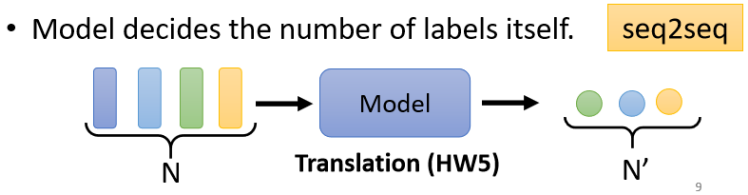
1. 输出个数与输入向量个数相同，即每一个向量都有对应的一个label或value（如命名实体识别NER、词性标注POS tagging等任务），也叫Sequence Labeling；



2. 只需要输出一个Label或value（比如文本分类、情感分析）；

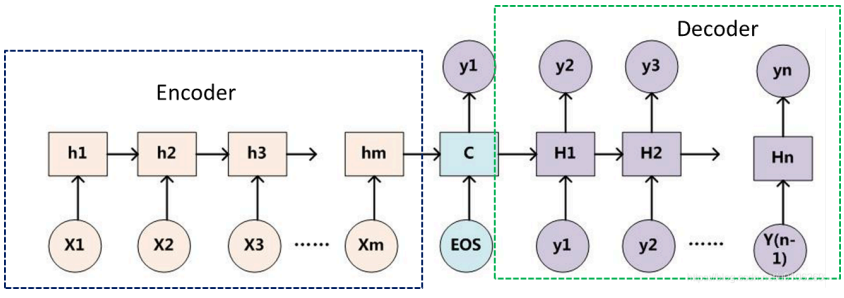


3 输出个数与输入向量个数不一定相同，机器要自己决定应该要输出多少个Label或value（比如文本翻译、语音识别），也叫做**sequence to sequence（Seq2Seq）**的任务。



Seq2Seq模型

Sequence-to-sequence (seq2seq) 模型，其输入是一个序列，输出也是一个序列。**其最重要的地方在于输入序列和输出序列的长度是可变的。 最基础的Seq2Seq模型**包含了三个部分，即Encoder、Decoder以及连接两者的中间状态向量，Encoder通过学习输入，将其编码成一个固定大小的状态向量C，继而将C传给Decoder，Decoder再通过对状态向量C的学习来进行输出。下图中的矩形（$h_1, h2, \ldots hm; H_1, H_2, \ldots . H_n$）代表了RNN单元，通常是LSTM或者GRU。



Seq2Seq模型与Encoder-Decoder框架的关系

> Seq2Seq 使用的具体方法基本都属于Encoder-Decoder 模型（强调方法）的范畴，Seq2Seq（强调目的）不特指具体方法，满足"输入序列、输出序列"的目的，都可以统称为 Seq2Seq 模型。

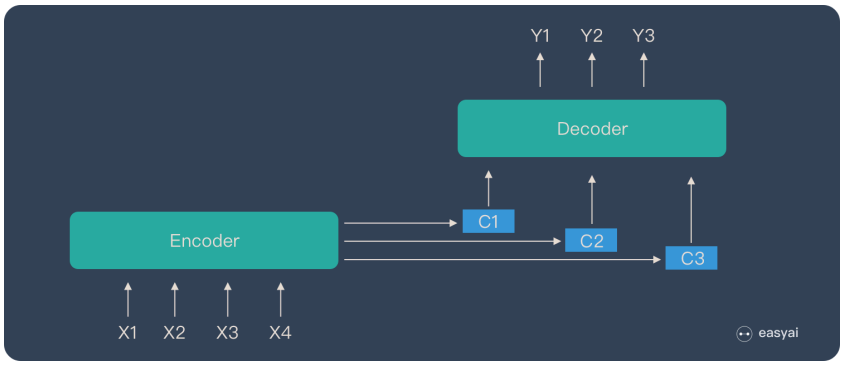
也就是说，Seq2Seq模型是基于Encoder-Decoder框架设计的，用于解决序列到序列问题的模型。

Seq2Seq模型的缺点

Seq2Seq模型缺点包括了RNN模块存在的缺点，和基础Encoder-Decoder框架存在的问题：

1. 中间语义向量C无法完全表达整个输入序列的信息；
2. 中间语义向量C对$y_1, y_2 \dots y_{n-1}$所产生的贡献都是一样的；
3. 随着输入信息长度的增加，先前编码好的信息会被后来的信息覆盖，丢失很多信息。

为了解决Seq2Seq模型的缺陷，引入了Attention机制，不再将整个输入序列编码为固定长度的"中间向量 C"，而是编码成一个向量的序列 $C_1, C_2, \dots$，如下图所示。将在下篇总结一下Attention机制。

> 
>

PS.李宏毅老师2021年的春季课程中已经没有将RNN完整的学习内容了（只作为补充资料），原话为："recurrent neural network 的角色，很大一部分都可以用 Self-attention 来取代了"。

参考文献

(强推)李宏毅2021春机器学习课程

李宏毅老师机器学习课程笔记

Encoder-Decoder综述理解(推荐)

从Encoder-Decoder(Seq2Seq)理解Attention的本质

Encoder-Decoder 和 Seq2Seq