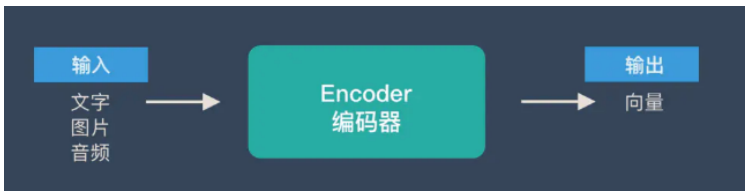
Encoder–Decoder 框架

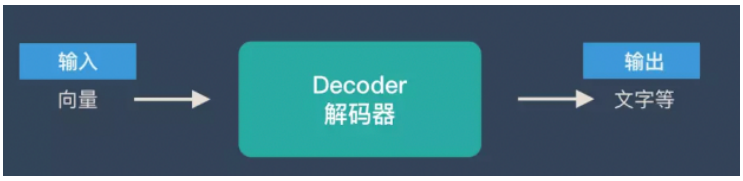
1. 概念

- Encoder-Decoder 框架可以看作是一种深度学习领域的研究模式，它并不特指某种具体的算法，而是一类算法的统称。
- 这个框架很好的诠释了机器学习的**核心思路**，即：将现实问题转化为一类可优化或者可求解的数学问题，利用相应的算法来实现这一数学问题的求解，然后再应用到现实问题中，从而解决了现实问题。
- 分为 Encoder 和 Decoder 两部分：

a) Encoder：将现实问题转化为数学问题



b) Decoder：求解数学问题，并转化为现实世界的解决方案



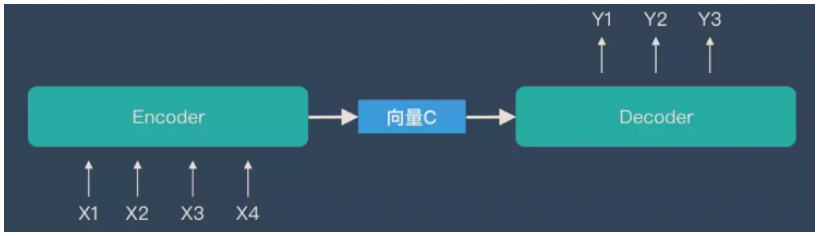
c) 合并图解

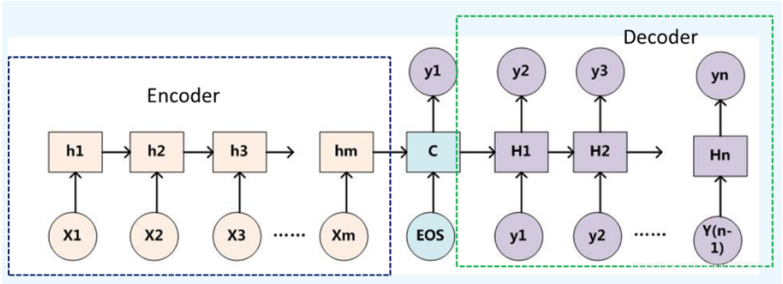
对于文本领域的 Encoder-Decoder 框架的实际模型案例比如，输入一个句子序列x_1, x_2, x_3 …，经过 encoder 进行非线性编码，获得一个中间向量C(中间语义)，decoder 根据这个向量和之前生成的历史信息去生成另外一个句子y_1, y_2, y_3 …。

$$C = f(x_1, x_2, x_3, \dots, x_n)$$

注：y_i除了受中间向量C的影响外，通常还受前序逐步生成的历史信息影响，即:

$$y_i = g(C, y_1, y_2, y_3, \dots, y_{i-1})$$





2. Encoder-Decoder 框架特点

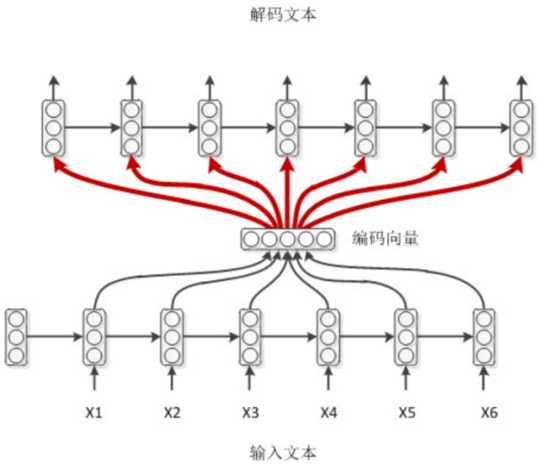
- 它是一个 end-to-end 的学习算法。
- 不论输入和输出的长度是什么，中间向量C的长度都是固定的(导致存在信息缺失问题)。
- 根据不同的任务可以选择不同的编码器和解码器(可以是 CNN、RNN、 LSTM、 GRU 等)，即可以任意选取不同的组合。
- 缺陷:

  基础 Encoder-Decoder 框架存在的最大问题在于信息缺失。

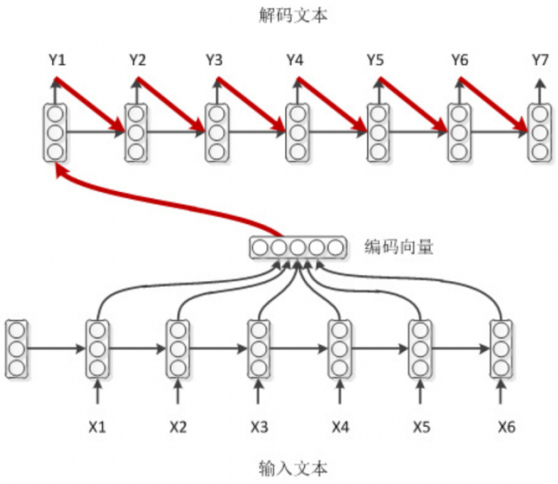
  Encoder 将输入编码为固定大小的向量的过程是一个"信息有损的压缩过程"，信息量越大，转化得到的固定向量中信息的损失就越大，这就得 Decoder 无法直接无关注输入信息的更多细节。输入的序列过长，先输入的内容携带的信息可能会被后输入的信息稀释掉或被覆盖了，那么解码的时候一开始就没有获得输入序列足够的信息，可能会导致模型效果比较差。

3. 不同 Encoder-Decoder 模式

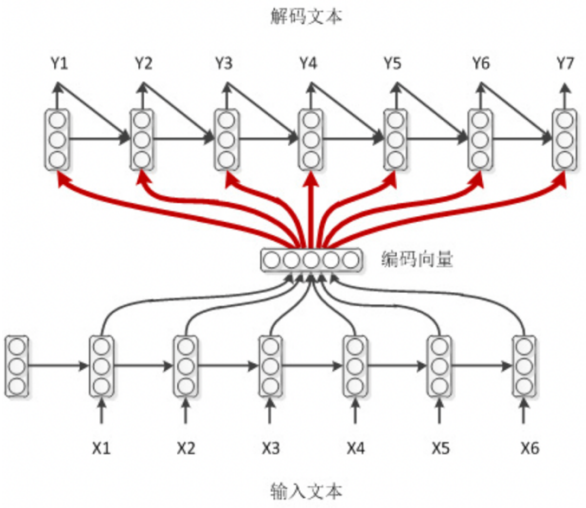
- 最简单的解码模式



- 带输出回馈的解码模式



- 带编码向量的解码模式



- 带注意力的解码模式