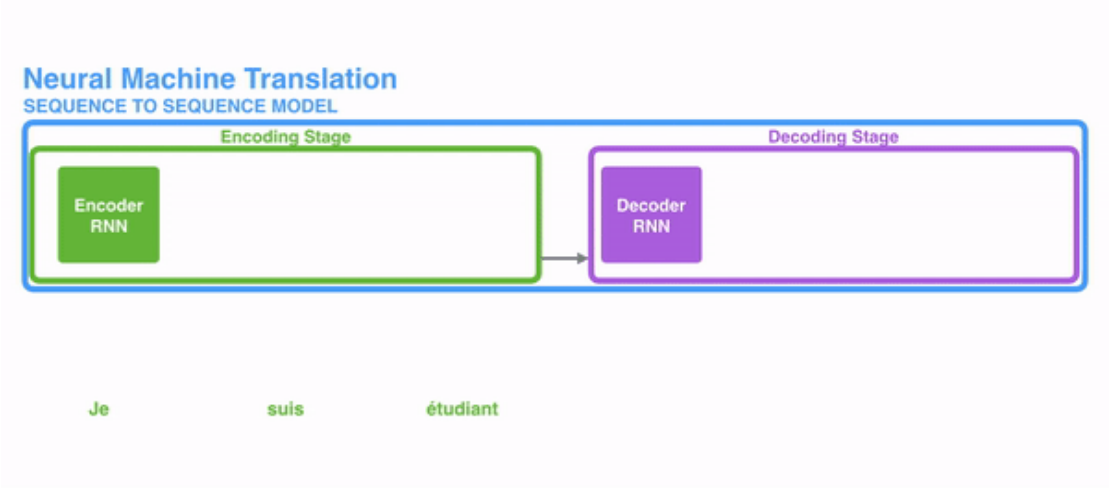
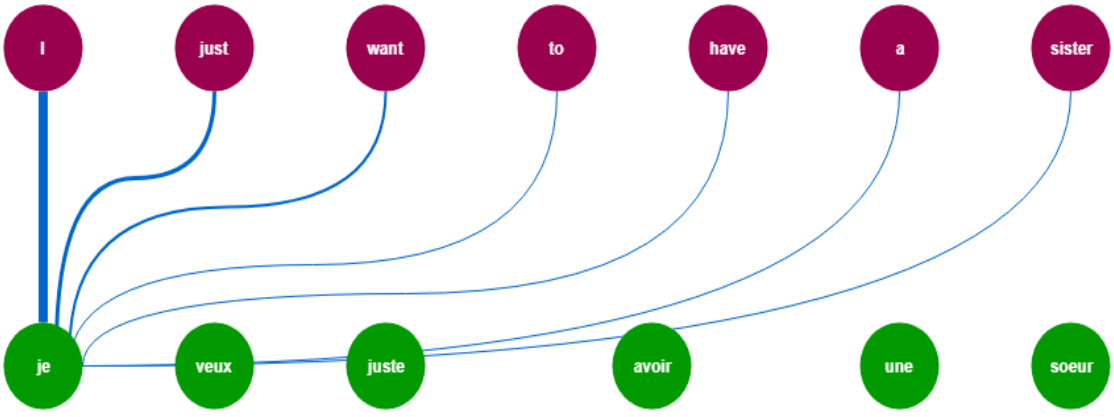
什么是注意力机制

深度学习发展起来以后，序列到序列建模成为很多任务的基本方案，如机器翻译、语音识别、语音合成、问答和对话等。序列到序列建模将输入序列用编码器压缩成一个固定维度的向量，再由一个解码器生成目标序列。如图所示。



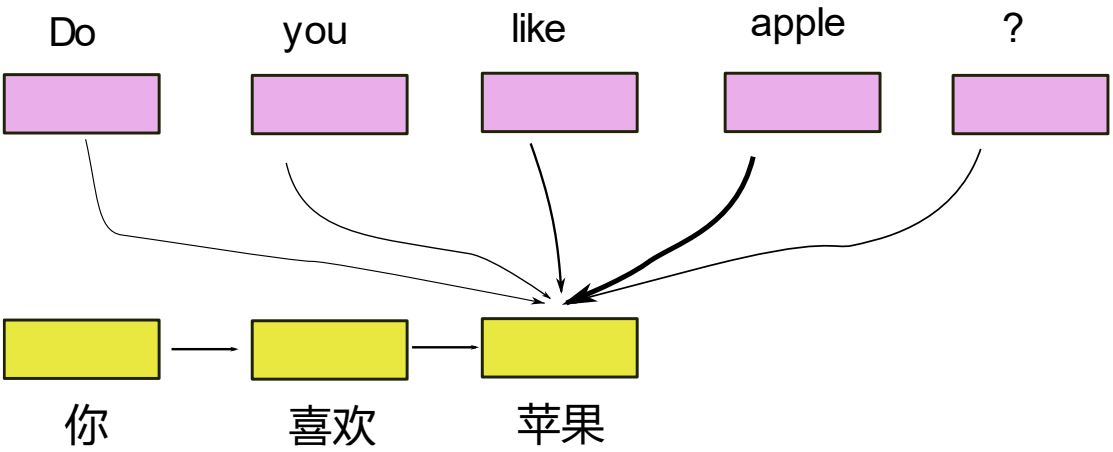
序列到序列模型虽然简洁，但有一个很大的问题，就是当输入序列很长时，往往很难生成合理的目标序列。这也很好理解，毕竟编码器给出的向量是固定维度的，当输入序列过长时，信息很容易丢失。这有点儿像人类的翻译员，如果听到的句子过长，可能在翻译时忘掉一些事情，导致漏译。

怎么办呢？研究者想出了一种注意力机制来解决这个问题。在这种新模型里，不再将输入序列压缩成一个向量，而是保留这一序列中每个元素的编码。在解码器时，以输入序列的编码为参考，逐渐生成目标序列，在生成过程中，不同时刻关注输入序列的不同位置，如下图所示。



以一个英中翻译模型为列，我们的目标是把“Do you like apple”翻译成中文。当前的状态是已经把“Do you like”翻译成了“你喜欢”。基于当前的状态，模型将注意力集中到了“apple”，并生成下一个词“苹果”。这样一点点转移注意力，最终把”Do you like apple”中所包含的所有意义表达出来。

注意力机制有点像带着纸笔的翻译员，在听到英语的同时把主要意思记录下来，然后按记下的内容一点点翻译成汉语，这样可以保证不会忘掉一些语义。



注意力机制是目前深度学习领域最重要的概念之一，在很多任务中都有应用。如图所示的一个给图片加标题的例子，当生成“Dog“这个词时，模型自动关注到了图片上趴在床下的狗狗。可见，注意力机制不仅可以帮助我们保留和利用语义，还可以为模型的行为提供某种解释，以帮助我们判断这一模型是否可信、可用。