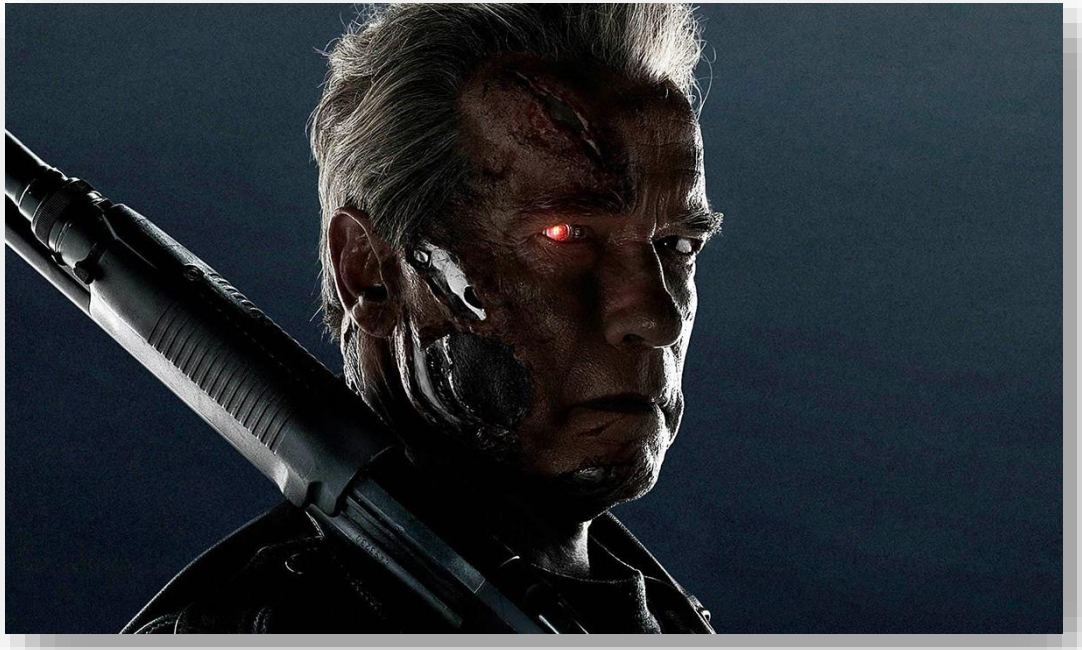
基于Seq2Seq注意力模型实现聊天机器人

目录

机器人

机器人的概念

- 具备机器和人的属性/特点
- 能思考、能行动、自动化、某些能力超强（如计算、推理、存储）
- 将人类从低端的重复劳动或特定场景中解放出来，提高生产效率



机器人

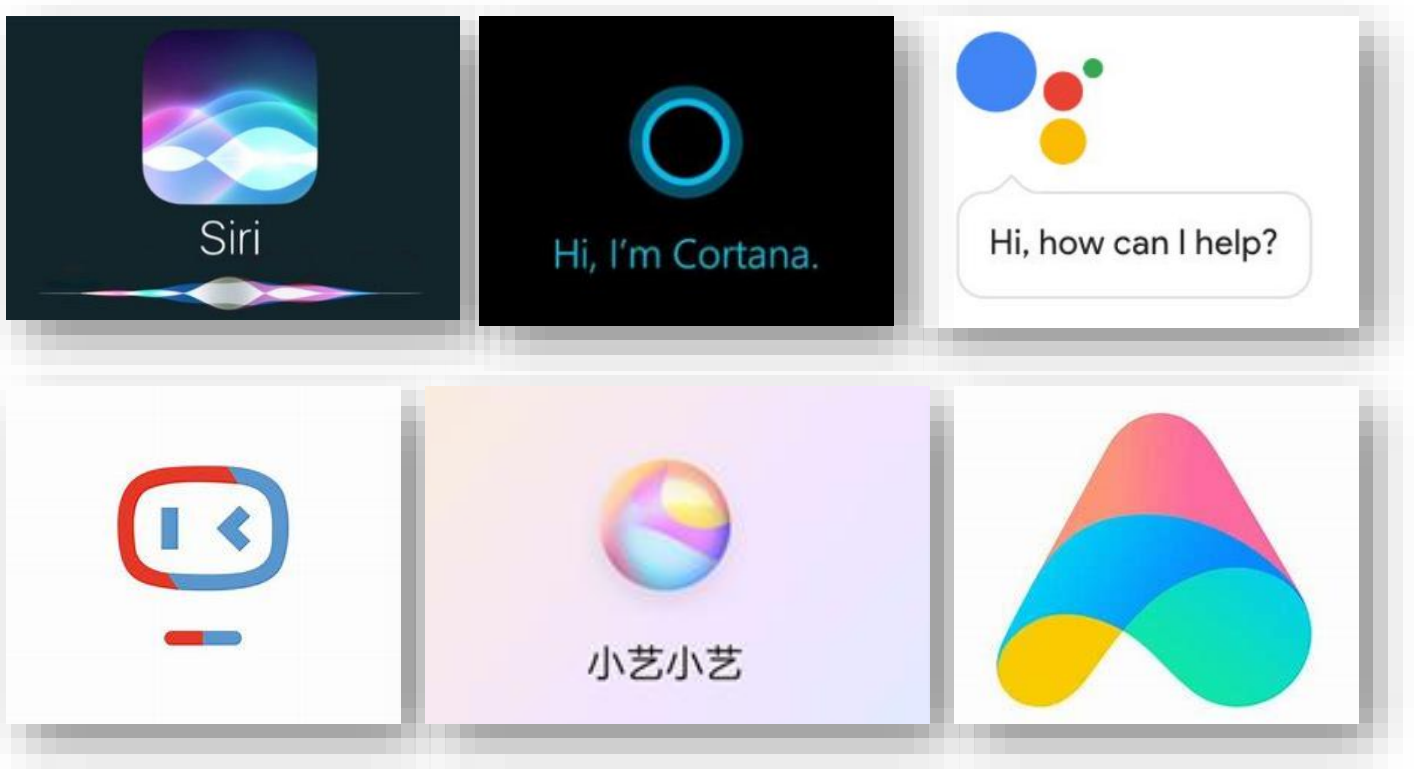
- 机器人（Robot）是一种可编程的多功能操作机；能用电脑改变或可编程的专门系统。
- 它既可以接受人类指挥，又可以运行预先编排的程序，也可以根据以人工智能技术制定的原则纲领行动。
- 人形机器人（Humanoid Robot）只是机器人定义中的一种。

机器人

应用

机器人

聊天机器人（Chatterbot）是能自动完成与人对话的计算机程序



机器人

ChatBot实现：Retrieval-based（知识库检索）

根据已有知识库进行问答匹配，但在知识库检索效率（深度优先、树形搜索）与精确度存在缺陷

机器人

ChatBot实现：Generative_Model(生成模型)

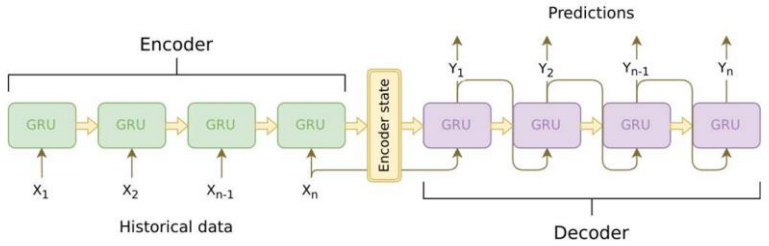
- 生成意料之外的答案
- 《**A Neural Conversational Model**》（**神经会话模型**）**是**最早应用于序列到序列框架建立对话模型的论文，结构不复杂，但效果是却很可观。(开启时代)
- 当前该领域的主流模型：RNN、LSTM、GRU……

案例目标

- 案例的主要目标：**搭建一个聊天机器人系统**
  - 了解语料库的基本处理，掌握一般自然语言处理的方法；
  - 理解循环神经网络及变体LSTM各个结构的作用，学会聊天机器人的实现方法。

本案例详细描述了一个聊天机器人模块实现的主要流程，从语料库预处理、基于注意力的seq2seq模型构建，再到模型训练及模型调用测试，复现了一个聊天机器人项目的详细过程。

分析方法及过程



语料库准备

语料库预处理

模型构建

模型训练

模型测试

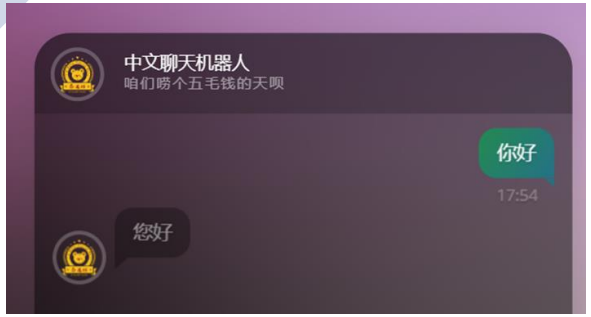
前端展示

all_dict.txt

source.txt

target.txt

```
Batch: 第 1000 次, 共2598
Epoch: 第 6 次, 共1000
Training Loss:  1.281  - Validation loss:  7.592
输入:  你走
target:  好吧。
Train输出:  轻轻地赶_UNK_UNK_UNK_UNK_UNK_UNK_UNK_UNK_UN
Inference输出:  轻轻地，我走了。。。。。。。。。。。
```



目录

语料库预处理

语料文件准备

- 语料库中存放的是在语言的实际使用中真实出现过的语言材料；语料库是以电子计算机为载体承载语言知识的基础资源；真实语料需要经过加工（分析和处理），才能成为有用的资源。
- 语料库（corpus，复数corpora）指经科学取样和加工的大规模电子文本库。
- 目前市面上已有的开源中文聊天语料有（https://gitee.com/chenyang918/chinese_chatbot_corpus）：
  - 豆瓣多轮
  - 电视剧对白语料
  - 贴吧论坛回帖语料
  - 微博语料
  - 小黄鸡语料

语料库预处理

语料文件准备

- 高质量的语料库应该具备以下要求：数据多、涵盖面广、专业等。
- 为确保构建的语料库质量，在本案例中使用的语料文件数量较少，便于控制，同时提高模型训练速度。
- 在工程内存储中文对话语料文件，语料文件以对话形式存储，指定存储的编码方式为UTF-8。

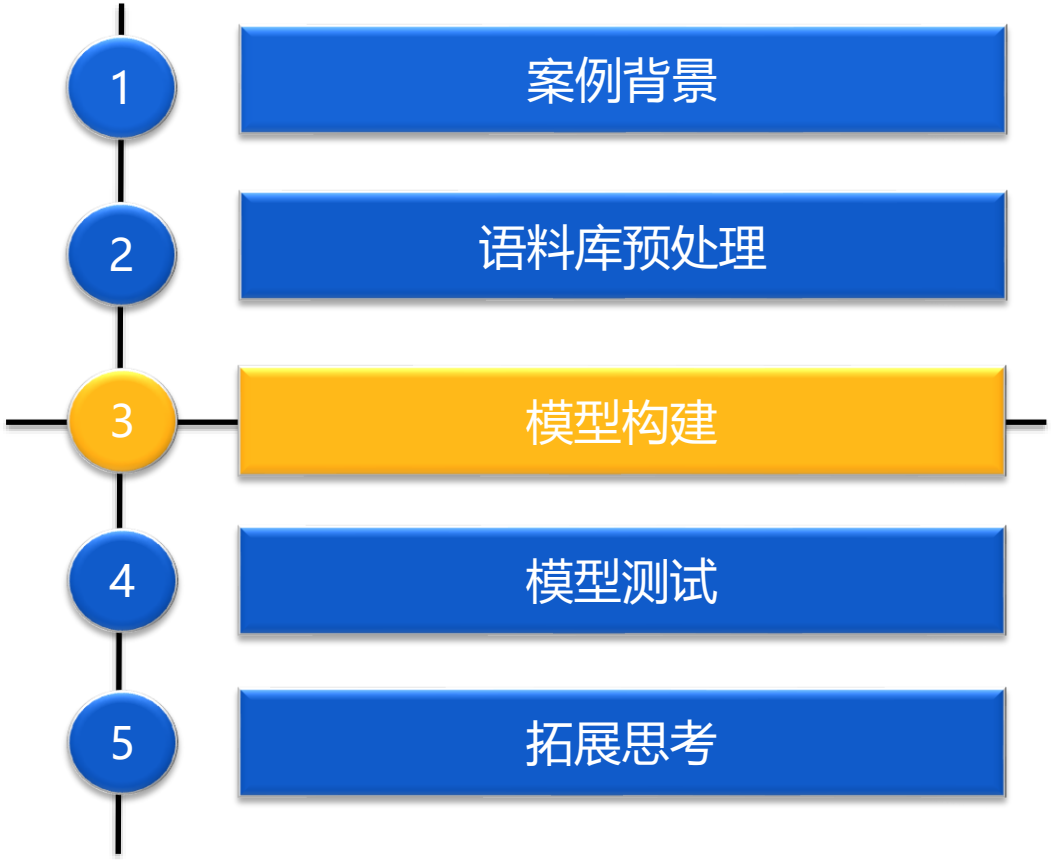
```
1 你好
2 您好
3 你吃了吗
4 我吃了呀
5 你吃的什么
6 我吃的三文鱼
```

语料库预处理

语料文件预处理

- 语料文件预处理包含以下步骤：
  - 语料读取：读取指定文件夹内的所有语料文件
  - 分词：对语料进行中文分词
  - 获取词典和问答数据
  - 文件保存：存储词典和问答语料到指定路径

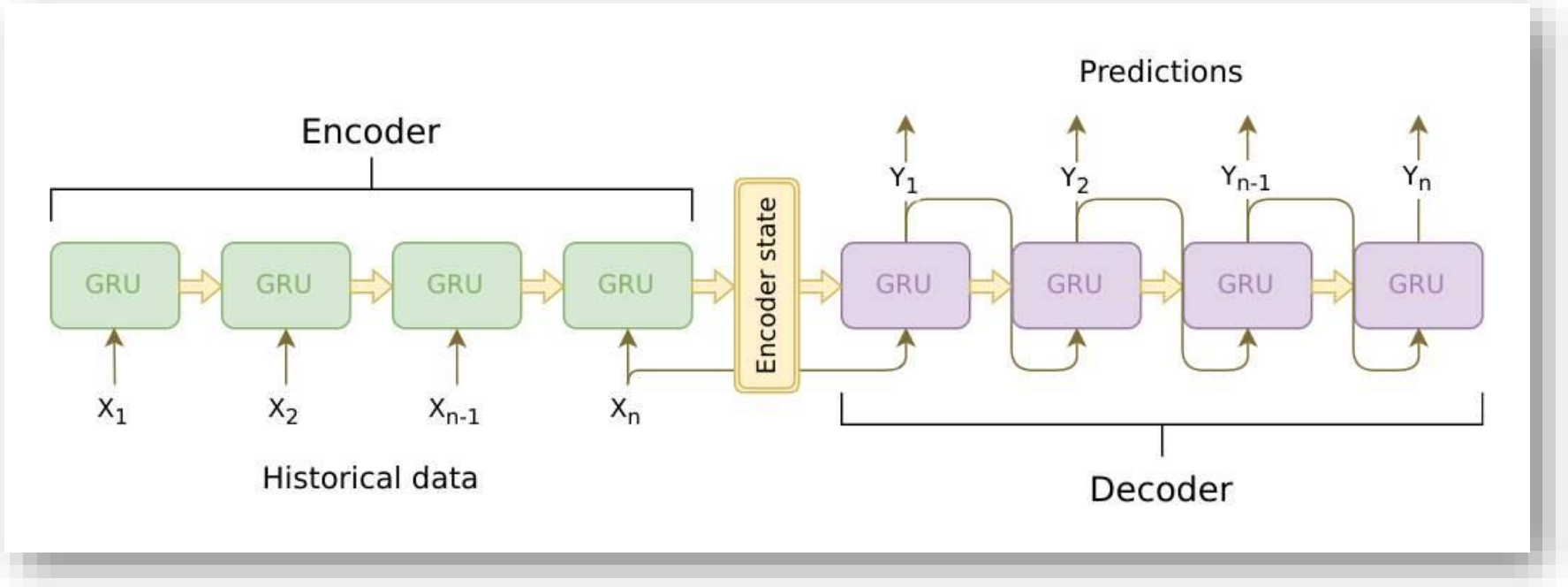
目录



模型构建

模型结构

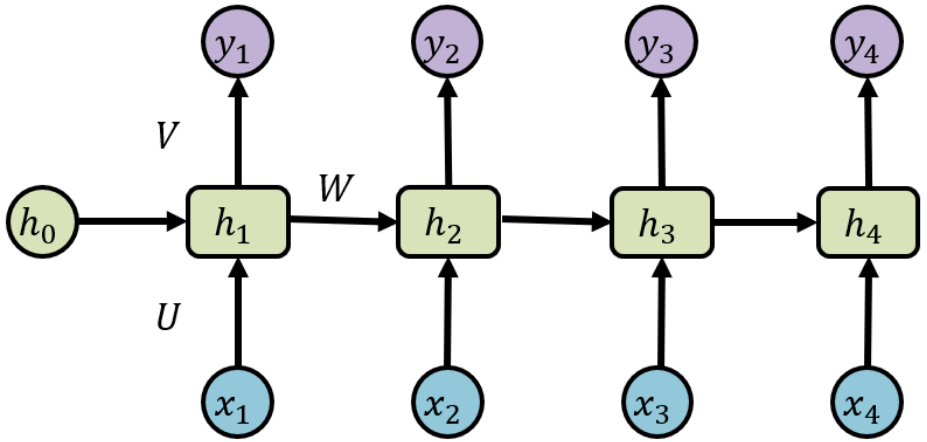
本案例中采用的模型主要是单层GRU构成的Seq2Seq模型，同时添加了BahdanauAttention机制



模型构建

GRU

- 循环神经网络（RNN）是一种用于处理序列数据的神经网络，比如某个单词的意思会因为上文提到的内容不同而有不同的含义， RNN就能够很好地解决这类问题。
- 但在长序列训练过程中RNN容易出现梯度消失和梯度爆炸问题，故提出相应变体，如LSTM、GRU等。

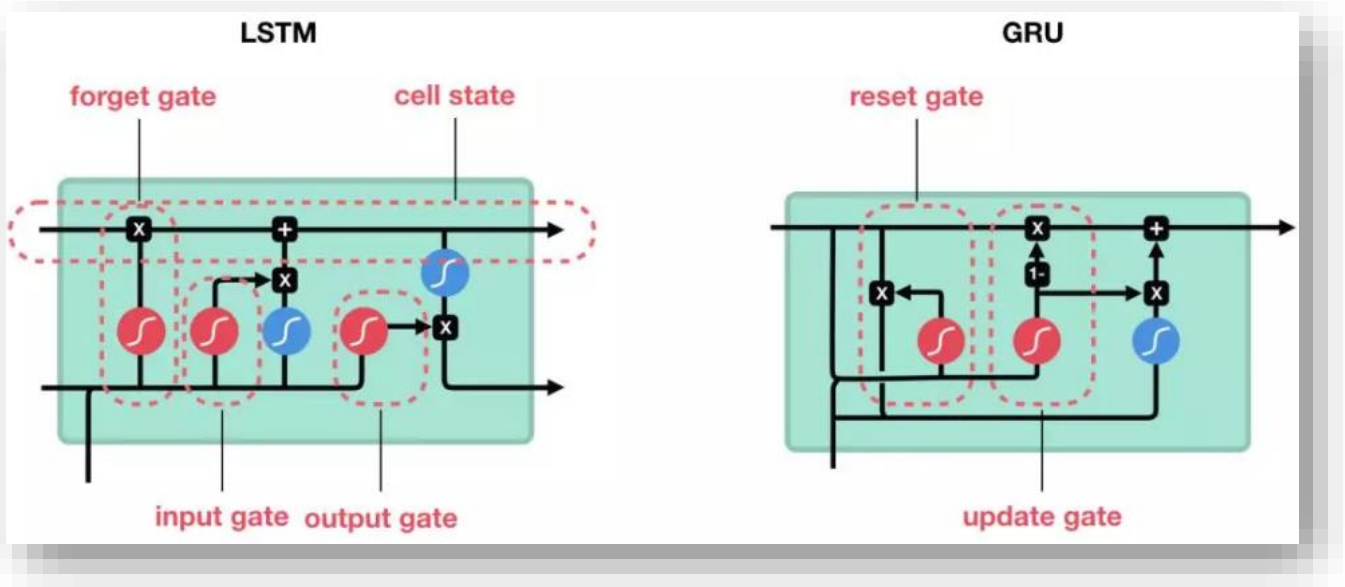


$$y_4 = f(Vh_4 + b)$$
$$= f(x_1, x_2, x_3, x_4)$$

模型构建

GRU

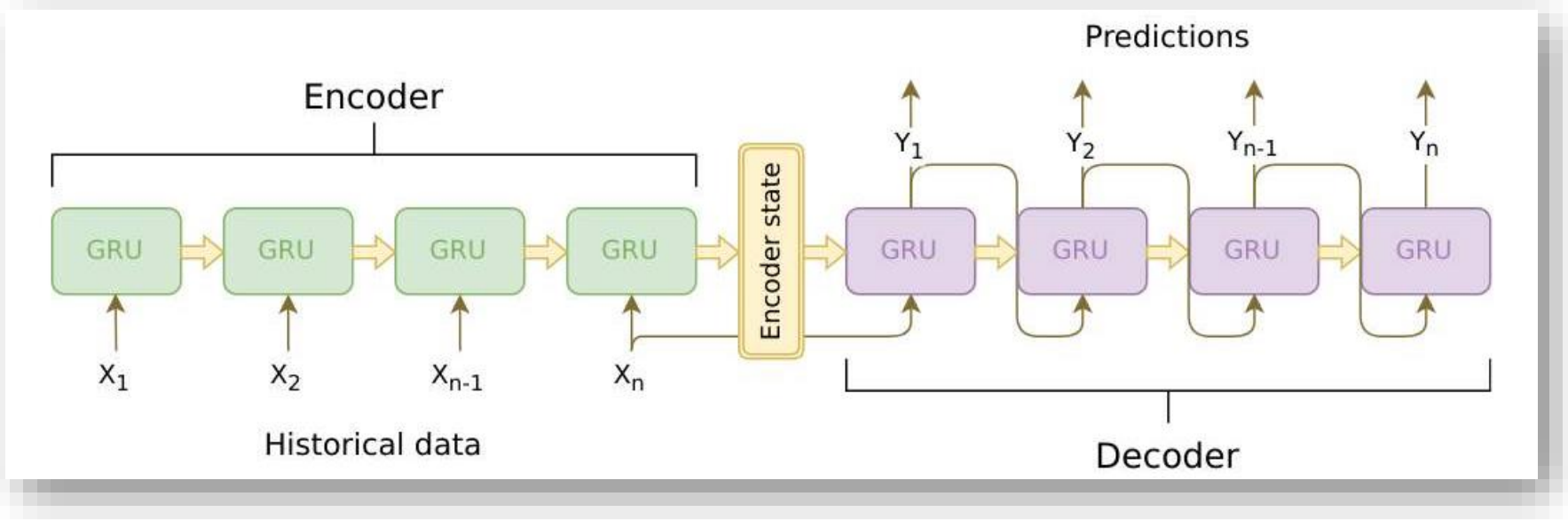
- GRU（Gate Recurrent Unit）是循环神经网络（Recurrent Neural Network，RNN）的一种。
- 相比LSTM，使用GRU能够达到相当效果的同时，还更容易训练，能提高训练效率。
- LSTM 和 GRU 是解决短时记忆问题的解决方案，它们具有称为“门”的内部机制，可以调节信息流。



模型结构

Seq2Seq结构

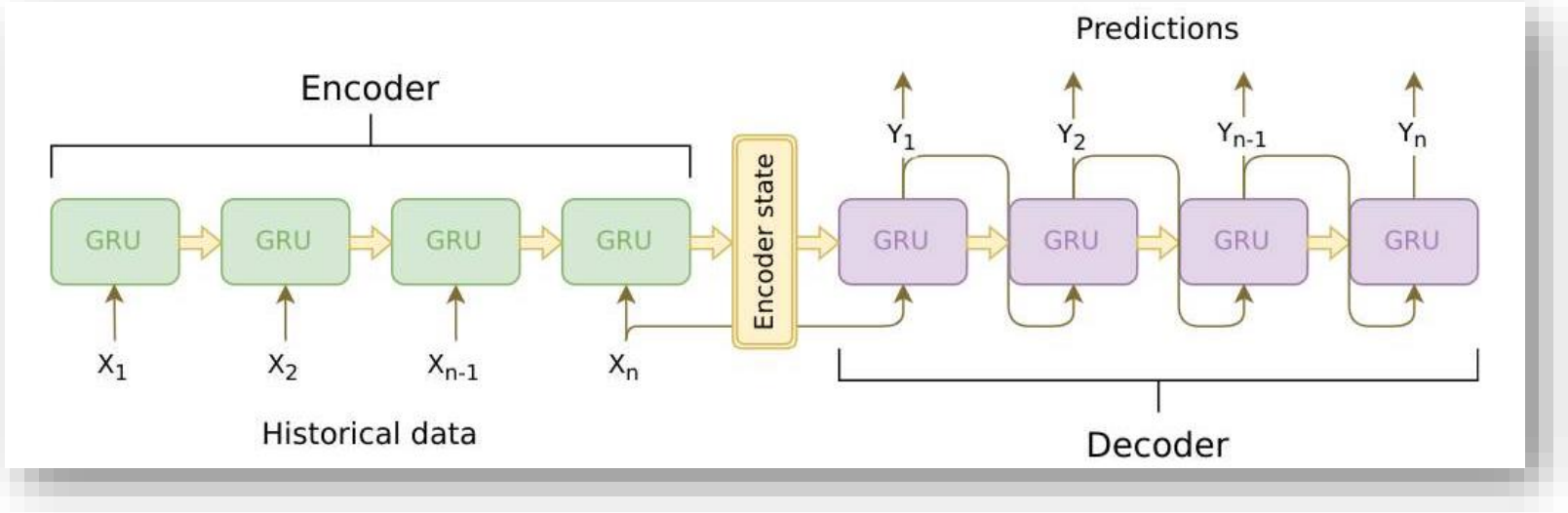
- Seq2Seq结构又叫Encoder-Decoder模型，也可以称之为Seq2Seq模型。
- 两个RNN网络构成，第一个RNN负责对输入数据编码，第二个RNN负责对编码后的数据解码。
- **Encoder结构先将输入数据编码成一个上下文向量c。**
- **Decoder结构负责对 c 进行解码。**



模型结构

Seq2Seq结构的应用

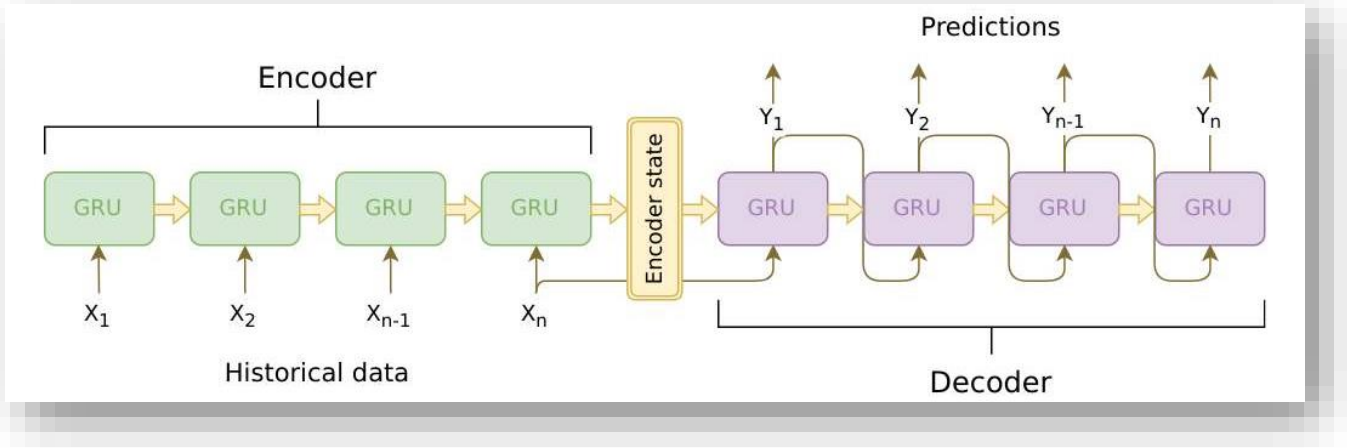
- 样本以<Source，Target>形式呈现
- 目标是给定输入句子Source，通过Encoder-Decoder框架生成目标句子Target。
- Source和Target可以是同一种语言，也可以不同，而Source和Target分别由各自的单词序列构成。



模型结构

Seq2Seq结构的应用

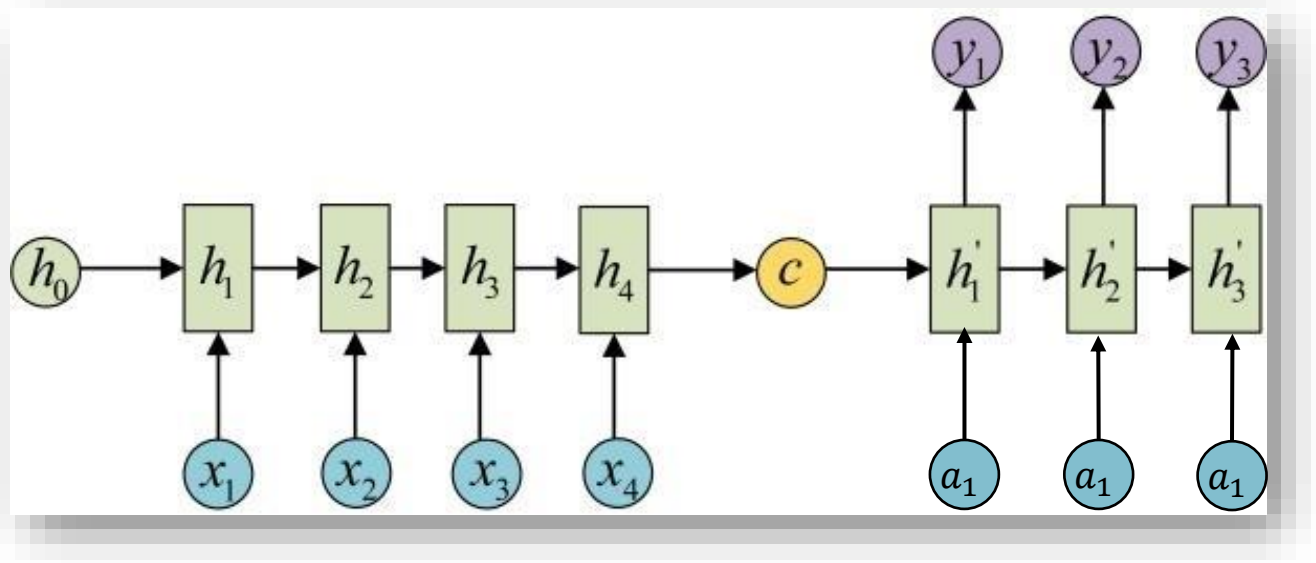
- **机器翻译**，Encoder-Decoder的最经典应用，事实上这一结构就是在机器翻译领域最先提出的。
- **文本摘要**，输入是一段文本序列，输出是这段文本序列的摘要序列。
- **阅读理解**，将输入的文章和问题分别编码，再对其进行解码得到问题的答案。
- **语音识别**，输入是语音信号序列，输出是文字序列。



模型结构

Attention机制

- 在普通Encoder-Decoder结构中，c的长度固定，对于长句子，其存储的信息可能不足，会造成精度下降。
- Attention机制通过在每个时间输入不同的 c 来解决这个问题。
- 关键思想：通过在对话时“关注”相关的Source的内容，在Target和Source之间建立直接的快捷连接。



模型结构

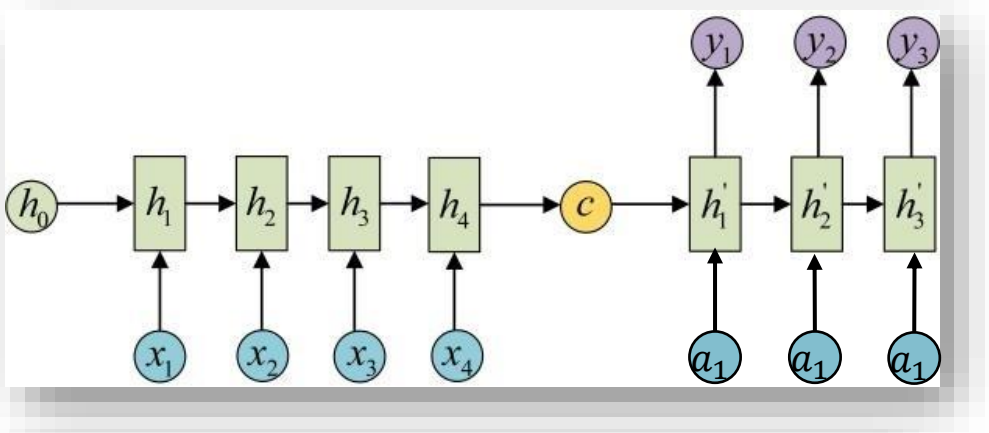
Attention机制（参考链接：https://zhuanlan.zhihu.com/p/91839581）

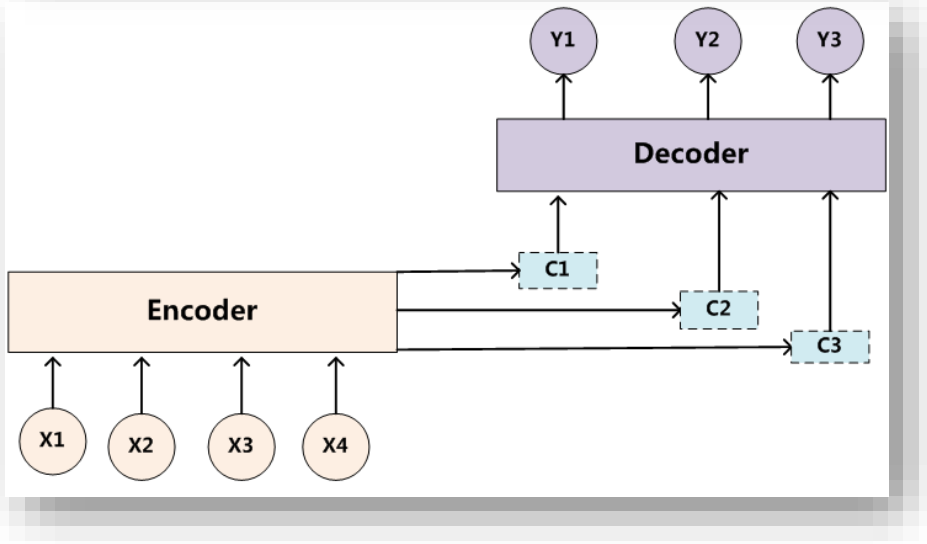
核心逻辑：**从关注全部到关注重点**。

将有限的注意力集中在重点信息上，从而节省资源，快速获得最有效的信息。

模型结构

对不同内容给予不同关注度





模型构建

数据准备

- 构造tensorflow支持的字典格式：加载词典并构建哈希表
- 添加开始和结束标记：读取分词后的语料、在前后添加“_BOS”、 “_EOS”并转换为ID向量
- 长度填充：填充“_PAD”统一长度

模型构建

模型构建

- H5保存的路径判断是否存在，若没有则新建
- GPU判断是否存在

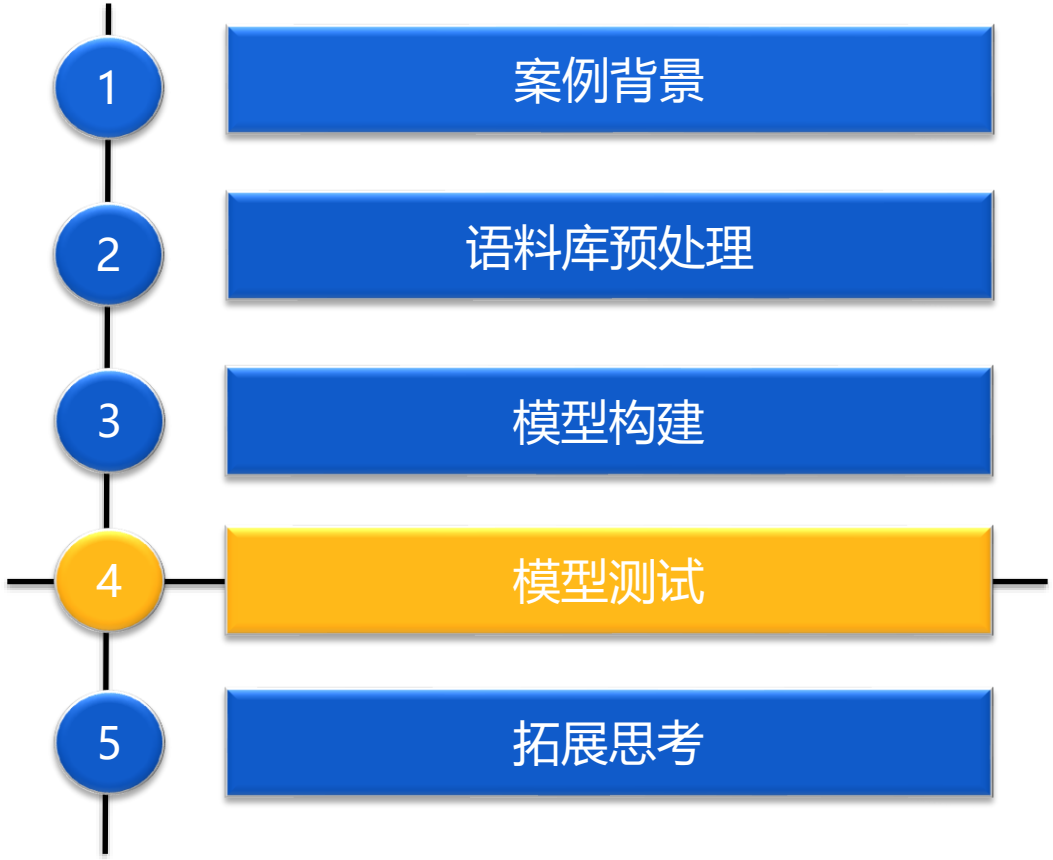
- Encoder端构建
- BahdanauAttention端构建
- Decoder端构建

模型构建

模型构建

- 模型编译：（Encoder → Decoder） → 优化器 →损失函数Loss function → h5模型参数保存
- 模型训练：训练步 → 迭代训练
- 模型测试：单个对话输入 → 分词、填充 → 编码 → 解码 → 预测的词id转为词
- 模型类化：训练和预测过程整理为class类

目录



模型测试

- 为方便演示，特实用Flask进行网页前端展示。

目录

问题

- 语料库
  - 高质量的中文对话的语料库较少
  - 电影对白：多人对话、情感分裂

- Seq2Seq模型
  - 模型训练速度
  - 有些问题没有标准答案（主观）

Thank you!