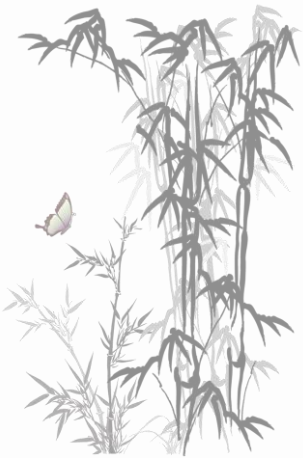
AI 模型生成诗歌对联

演示 概念 代码 分析

02/2022

“课程” 安排

- [15 mins] 代码模型分享和快速现场演示
- [15 mins] 简介两种深度生成语言模型原理
- [30 mins] 代码解读和讨论
- [15 mins] 生成文字的质量分析
- 更多演示和讨论



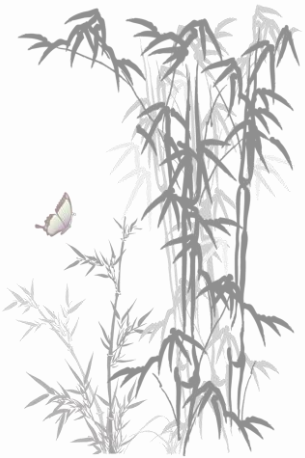
代码分享和现场演示

github.com/hululuzhu/chinese-ai-writing-share

- Inference目录：两个colab使用训练好的模型进行文字生成
- Training目录：两个colab可以训练你自己的模型

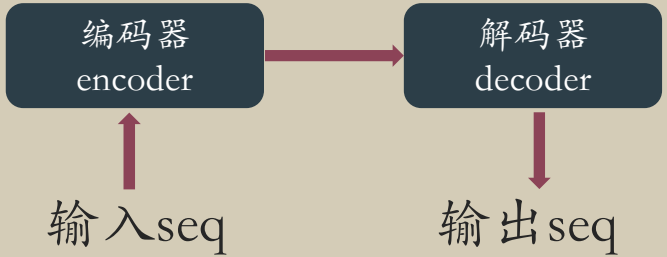
现场演示

- 请访问这个表格，填入你想模型生成的对联或诗歌
- 欢迎评论，表扬，吐槽。。。



AI写作模型

通用架构



Seq2Seq

实现方式

循环神经网络

简称RNN

输入作为sequence

简单有效

高级循环神经网络

LSTM
GRU
Bi-LSTM

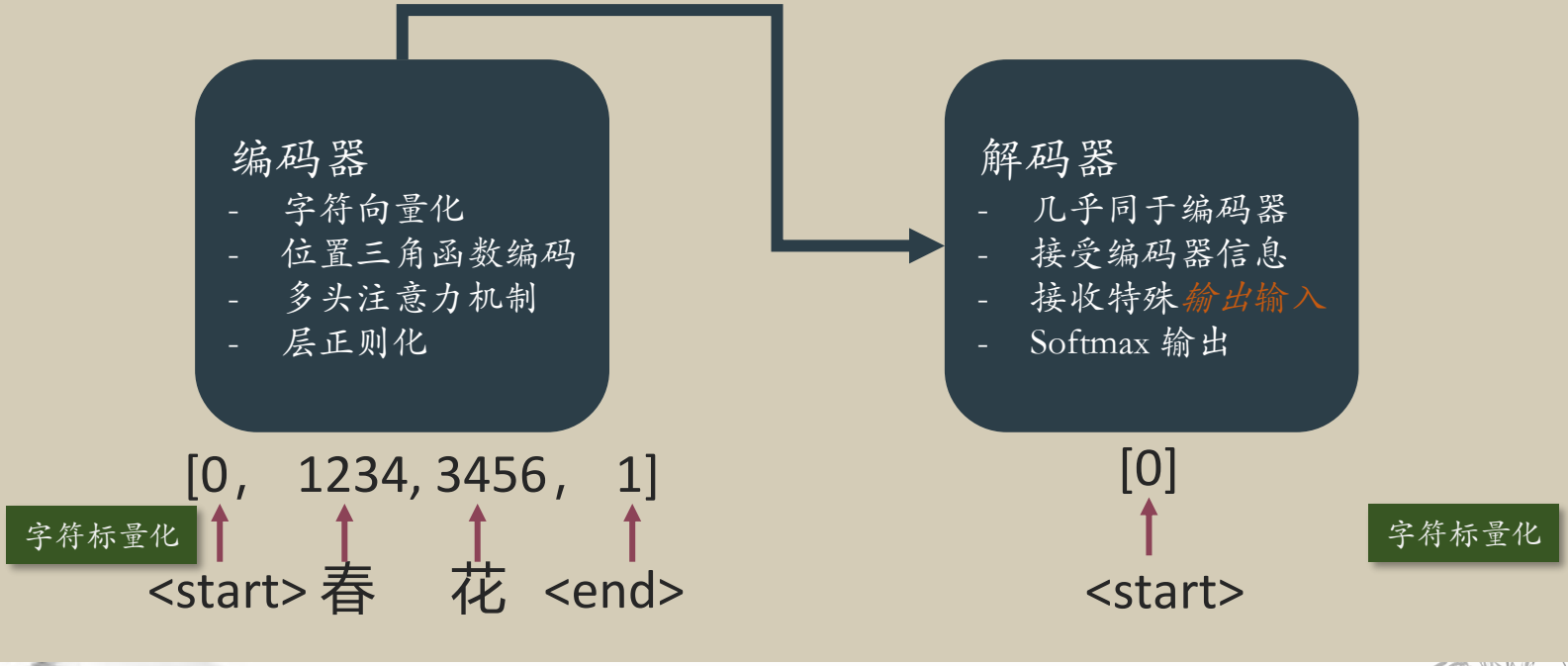
2018前比RNN更强的sequence处理能力

最强主流模型

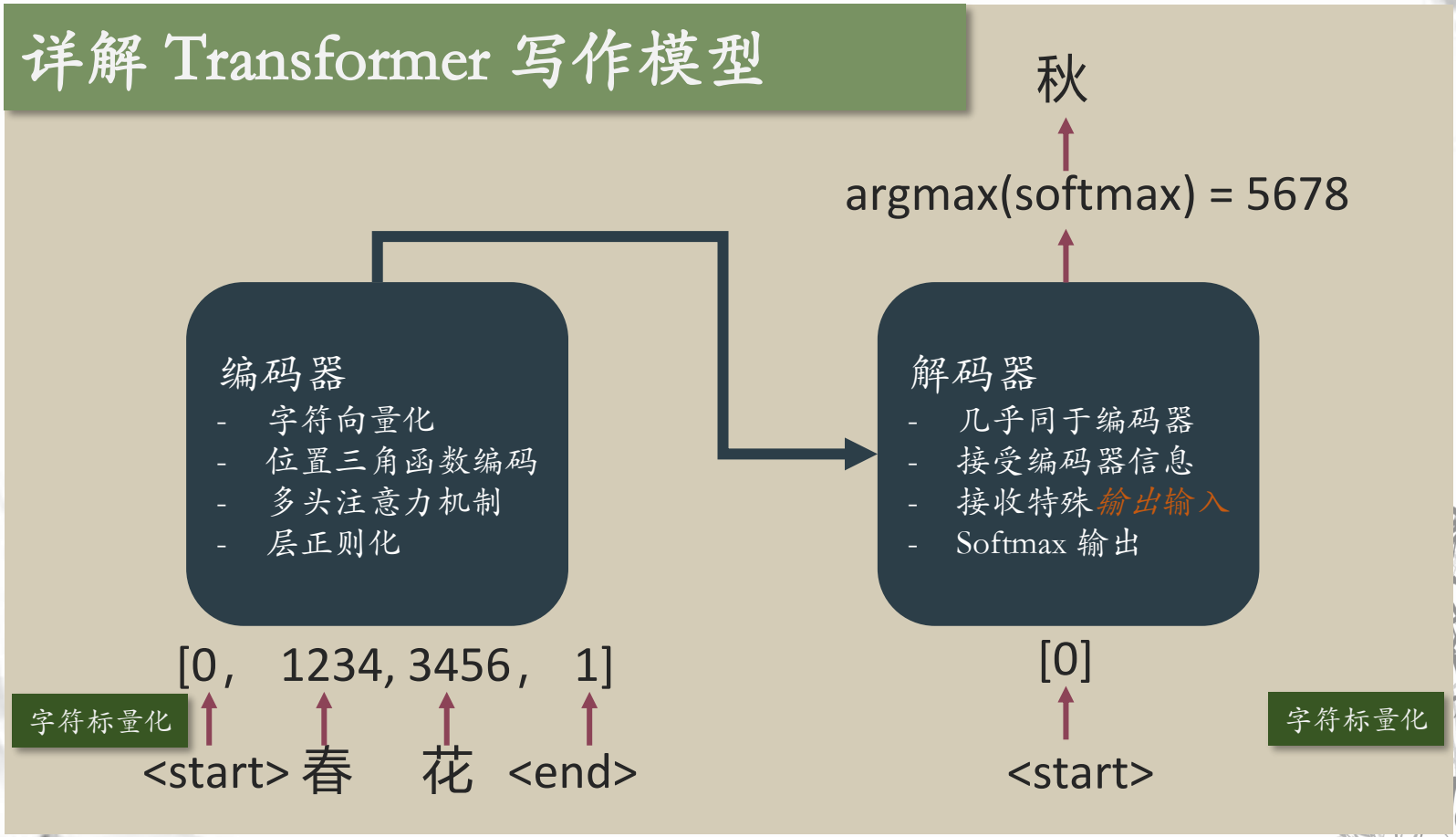
Transformer

2017 前主流用于翻译
2018 BERT爆发
2019开始 GPT
LAMDA MUM ‘悟道’ 都基于此

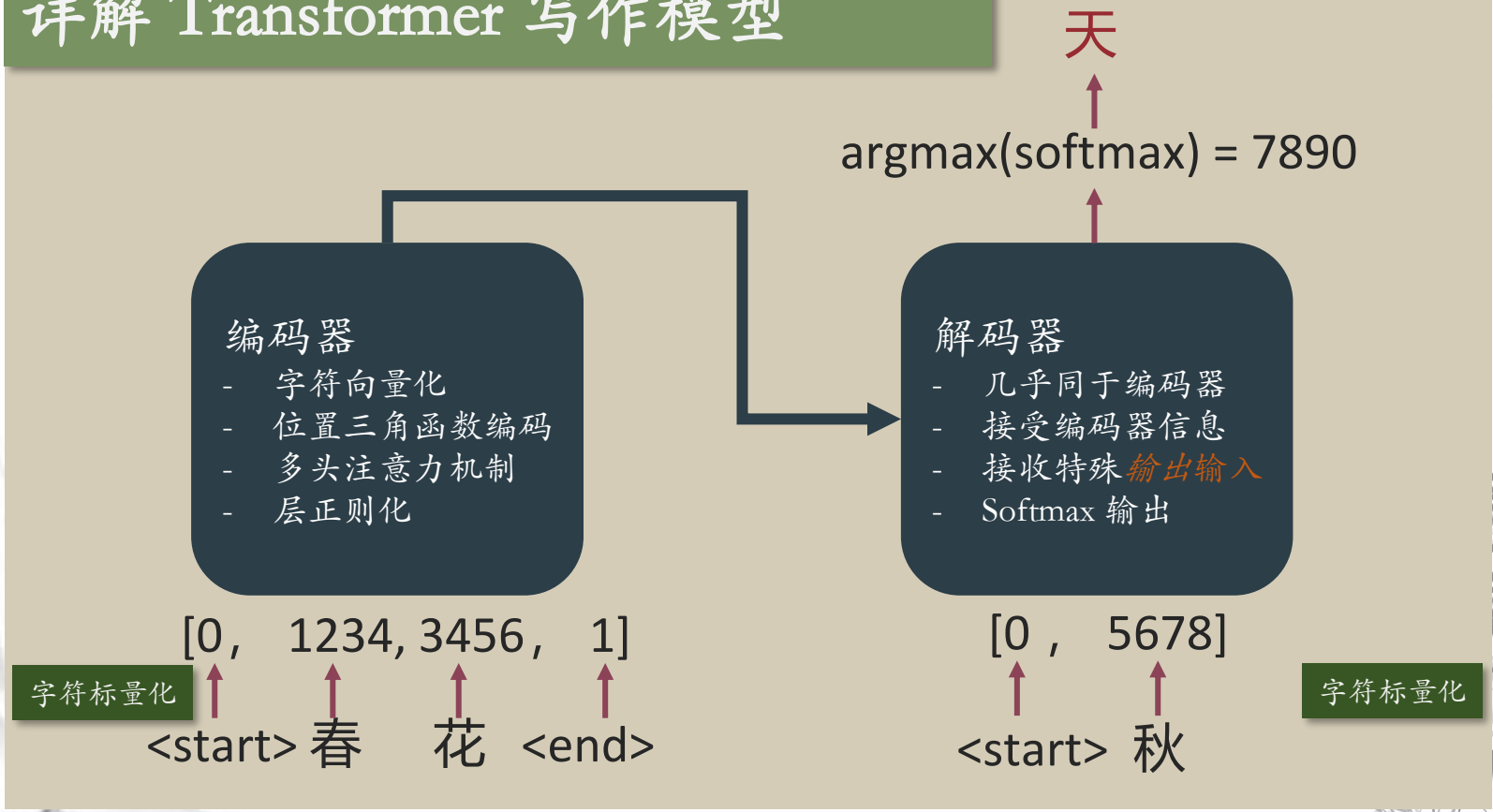
详解 Transformer 写作模型



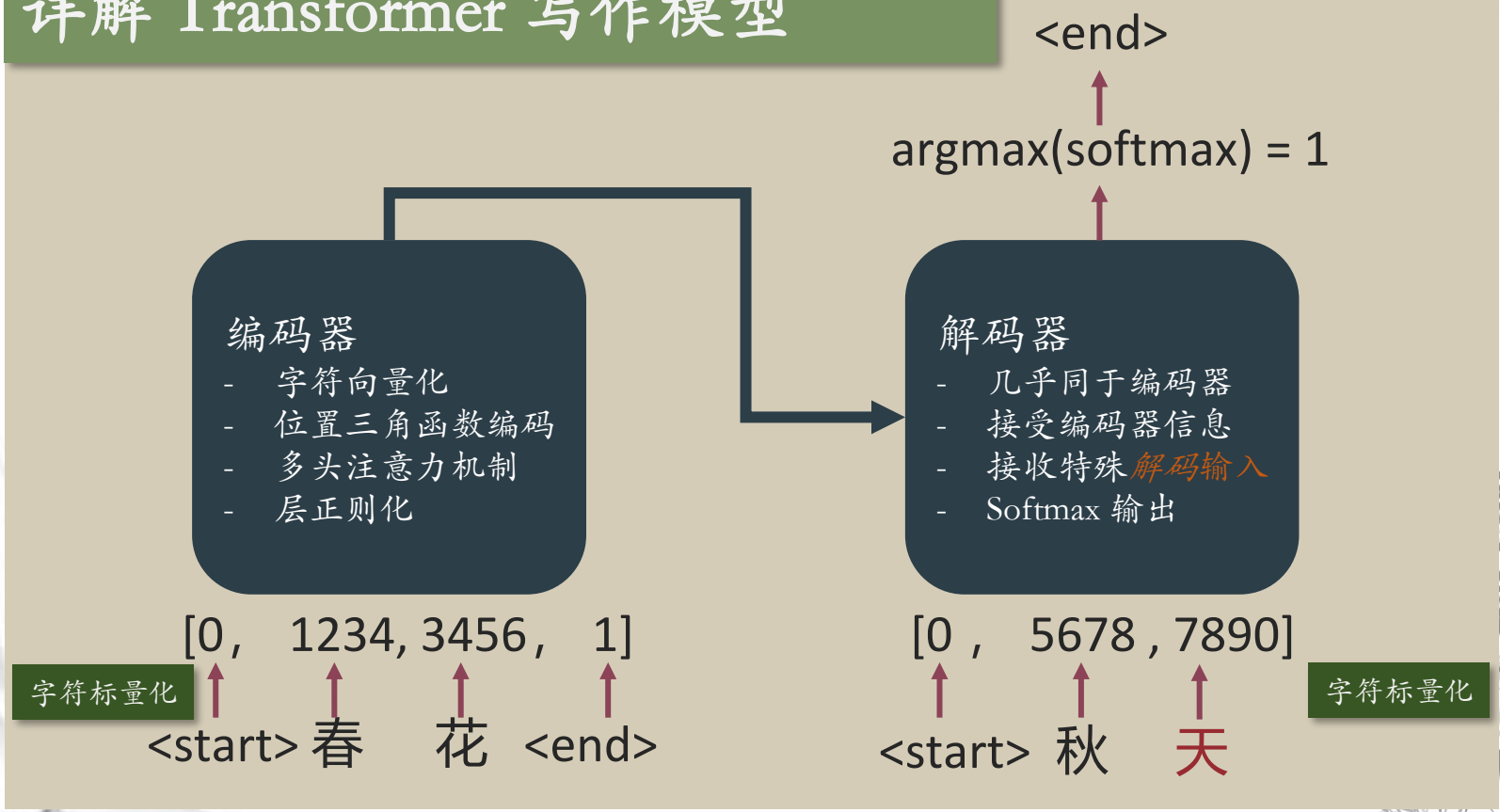
详解 Transformer 写作模型



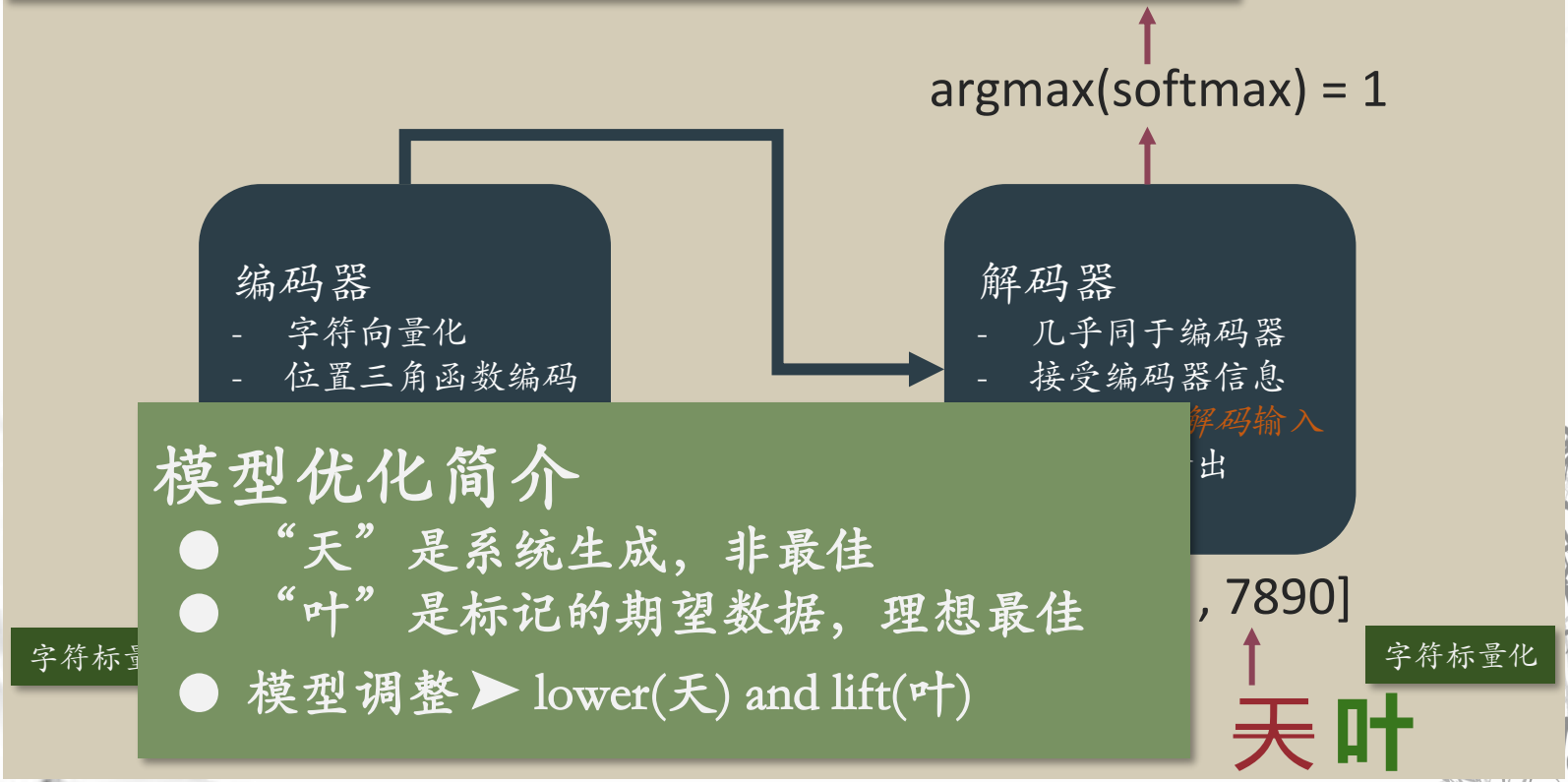
详解 Transformer 写作模型



详解 Transformer 写作模型



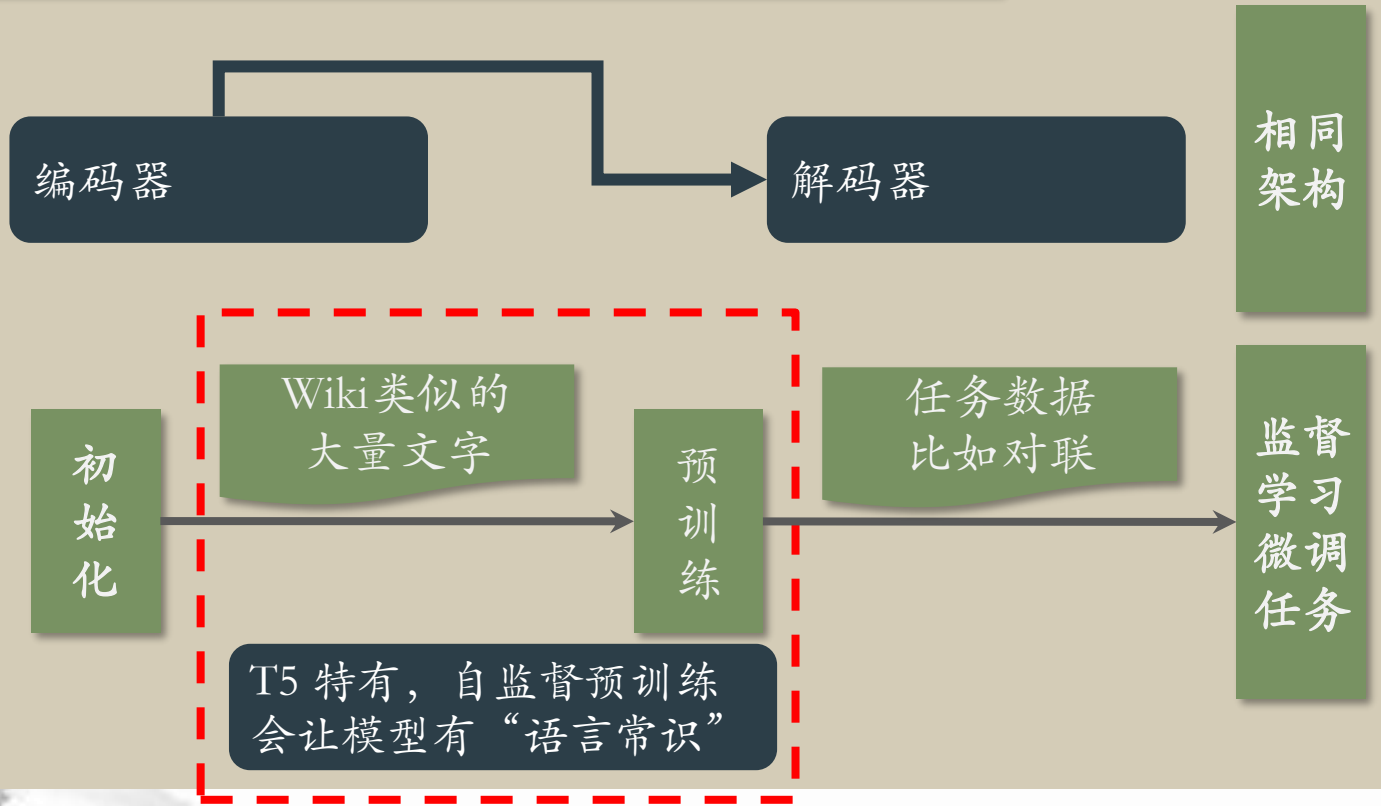
详解 Transformer 写作模型 - 训练优化



模型优化简介

- “天”是系统生成，非最佳
- “叶”是标记的期望数据，理想最佳
- 模型调整 ➤ lower(天) and lift(叶)

基于 T5 写作模型有何不同？



模型输出的常用算法设置

- Greedy算法，每个位置输出最佳token
- Beam Search算法，保持搜索树的最优 N 路径
- Sampling取样选项，是否根据比率取样
- TopK 取样，每个位置考虑最优的 K 个token
- TopP 取样，最优的n个token的比例总和超过 P
- Temprature 参数，锐化或钝化模型输出比例

注：通常 TopP 取样效果最佳，但有随机性，为固定演示效果，代码关闭了 sampling 使用 beam search

代码解读和讨论

github.com/hululuzhu/chinese-ai-writing-share

- Inference 目录
  - 两个colab使用训练好的模型进行文字生成
- Training 目录
  - 两个colab笔记本可以去训练你自己的模型

Colab页面代码解读ing

生成文字“赏析” - 对联

比较成功的例子

- 例1： 能够结合大环境背景

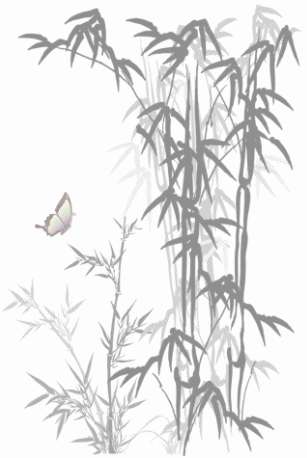
  上： 虎跃龙腾生紫气

  下： 莺歌燕舞庆新春 （T5）

- 例2：对仗工整，意思清晰

  上: 春雨丝丝润万物

  下: 和风阵阵暖千家



生成文字“赏析” - 对联

不太成功的例子

- 例1： 无对仗，词句重复

  上： 玉笛横处梅花落

  下： 玉笛横时柳絮飞

- 例2： 只追求对仗而导致语义不通顺

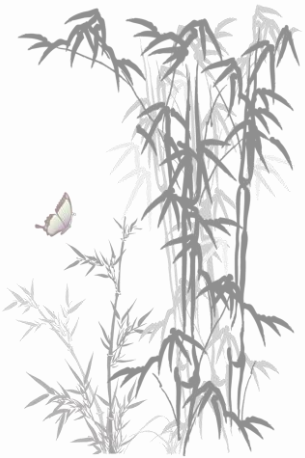
  上: 雾里看花花不语

  下: 风中听雨柳无声

- 例3: 对仗工整齐，语义清晰但是没有结合时代环境

  上: 虎跃龙腾生紫气

  下: 莺歌燕舞颂和风 （transformer）



生成文字“赏析”-诗歌

只给主题的诗歌：

好的例子：

- **主题：** 春日
  正文： 春风吹雨过，春色满江城。不见花开尽，空余柳色清
  优点： 景中有情，既写出了春景，又写出了对春光易逝的惋惜之情（后两句）。

- **主题：** 送别
  正文： 送君江上别，此别意何如。客路三千里，家山十二余。
  风烟随处好，云水向人疎。莫怪频回首，春深未得书。
  优点： 离别之情真挚，同时能寄情于景。

生成文字“赏析” - 诗歌

只给主题的诗歌：

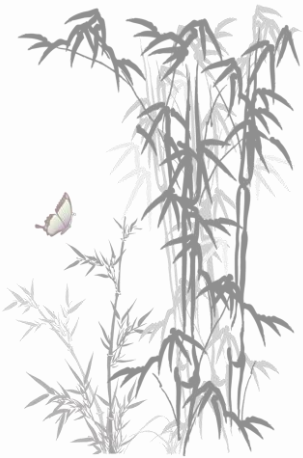
不好的例子：

- 主题：送别

  AI 正文：

  一别一年别，相逢一别离。相逢无别恨，不见有归期

  缺点：语句不太通顺，语义前后矛盾。



生成文字“赏析” - 诗歌

给主题以及仿写作者的诗歌：

主题：春日　**仿写作者：**王维

作者诗词特点：诗中有画，画中有诗。

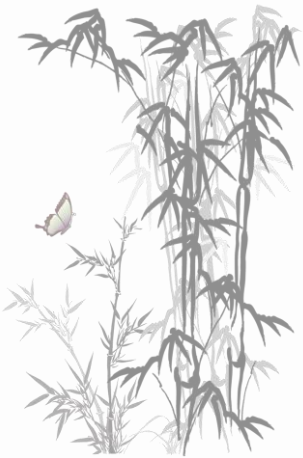
作者诗词代表《鸟鸣涧》：

人闲桂花落，夜静春山空。月出惊山鸟，时鸣春涧中。

AI 仿写：《春日 | 模仿 王维》

春日东郊外，春风北户前。鸟啼花落处，人语柳垂边。

优点： 画面感非常强，对仗工整。



生成文字“赏析” - 诗歌

给主题以及仿写作者的诗歌：

主题： 送别　**仿写作者：** 杜甫

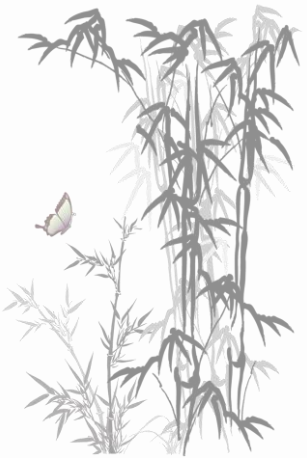
作者诗词特点： 沉郁顿挫，语言精炼，格律严谨，感情真挚，平实雅谈。

作者诗词代表《奉济驿重送严公四韵》：

远送从此别，青山空复情。几时杯重把？昨夜月同行。
列郡讴歌惜，三朝出入荣。江村独归处，寂寞养残生。

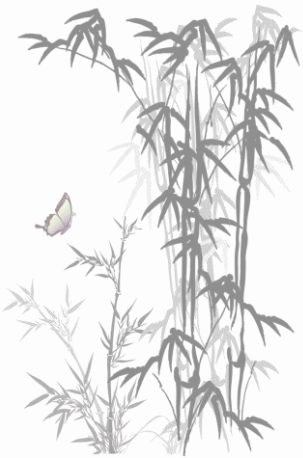
AI 仿写： 《送别 | 模仿 杜甫》

送君江上别，此别意何如。落日孤舟远，秋风一雁疎。
云山连楚塞，风雪满荆吴。莫作关中客，愁看鬓欲疏。



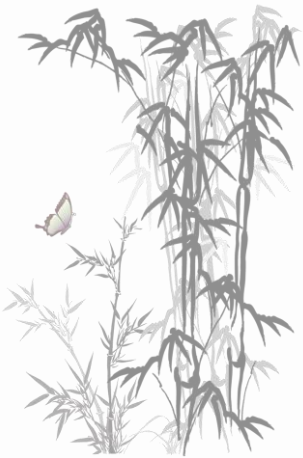
这个 seq2seq 架构（代码）还能做什么？

- 翻译，古文白话文翻译，命题写歌词或英文写作
- 写散文或者高考作文?
  - 通常512字符限制会有影响
  - 看图写作文需要多模态multi-modal
- 写[很长的]小说?
  - 效果很差，需要GPT这样的decoder only架构
- 写代码?
  - 有可能，但一般decoder only，比如codex
- 其他？欢迎讨论



引用

- 理论背景
  - Google: [1706.03762] Transformer Paper
  - Google: [1910.10683] T5 paper
  - HuggingFace: How to generate text
  - OpenAI: GPT3
  - 澜舟科技: [2110.06696] Mengzi Chinese T5 paper
- 数据
  - GitHub chinese-poetry: 最全中华古诗词数据库
  - GitHub wb14123: 70万条对联数据库
- 代码模型
  - 澜舟科技: 开源 mengzi-t5-base 模型
  - GitHub CyberZHG: keras-transformer
  - GitHub Shivanandroy: SimpleT5



感谢大家，欢迎讨论提问

AI 模型生成诗歌对联

演示 概念 代码 分析

02/2022