深度之眼
deepshare.net

专题二：实体识别挑战赛 编码器与解码器串讲

导师：Gauss

深度之眼
deepshare.net

目录

深度之眼
deepshare.net

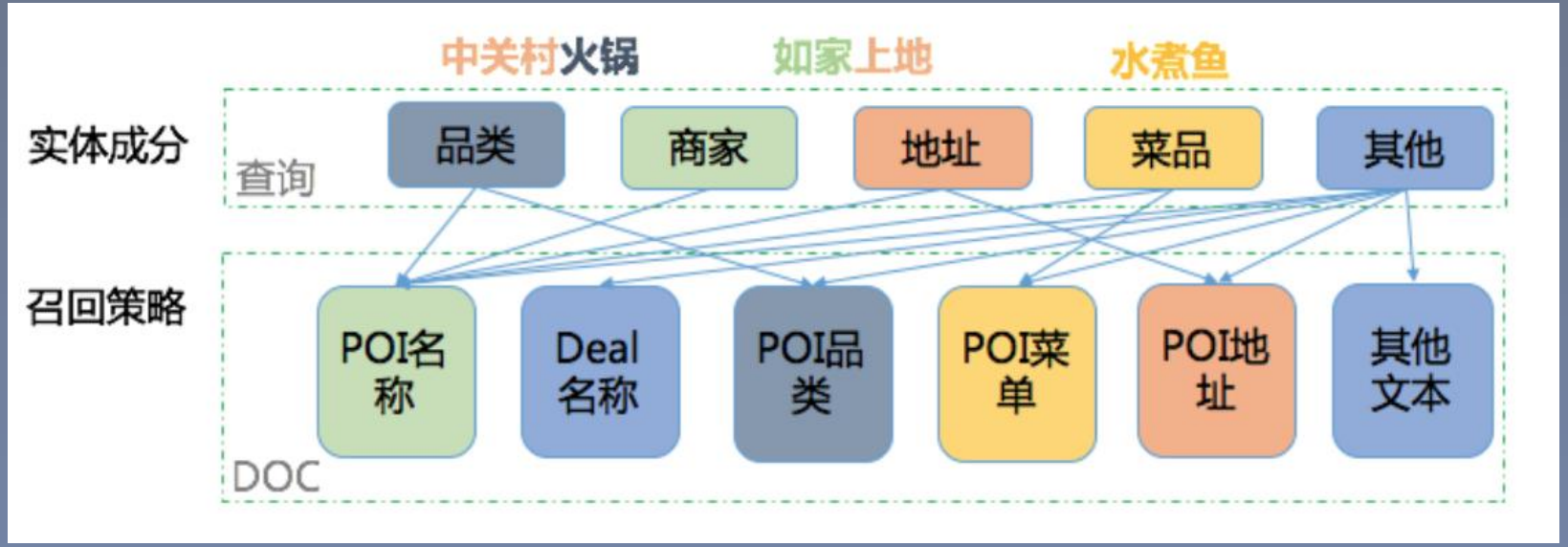
1、NER任务再谈

NER任务应用场景

应用场景:

搜索、知识图谱、客服机器人等

以搜索为例:



NER任务相关资料

- 美团NER的应用

https://tech.meituan.com/2020/07/23/ner-in-meituan-nlp.html

- 工业界如何解决NER问题？

https://zhuanlan.zhihu.com/p/152463745

- CS224N

https://zhuanlan.zhihu.com/p/61601575

深度之眼
deepshare.net

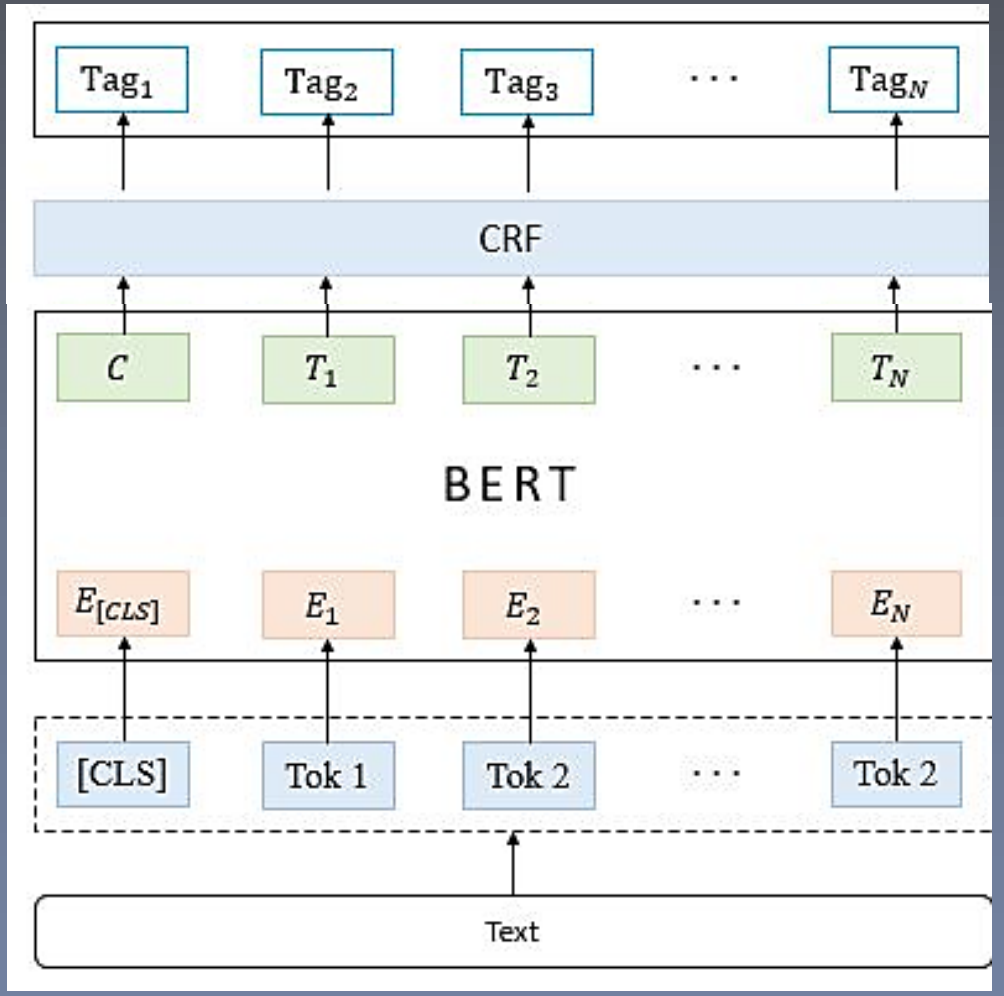
2、编码器串讲

编码器扩展

- BERT-CRF
- BERT-**BILSTM-CRF**
- BERT-**IDCNN-CRF**
  （https://arxiv.org/pdf/1702.02098.pdf）
- BERT多层表示的动态权重融合
- 不同预训练模型替代BERT模型(详见专题一)

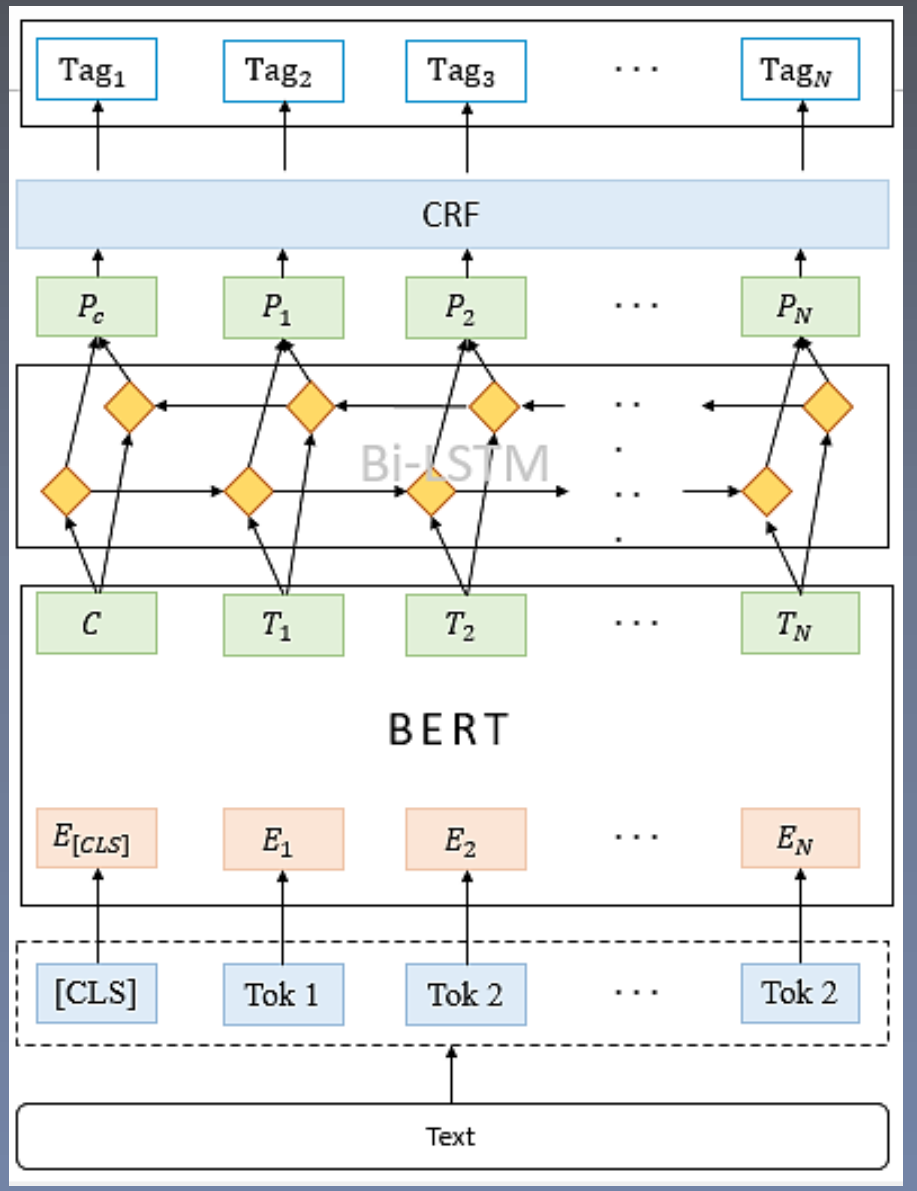
BERT-CRF

BERT-CRF对于在NER任务中目前是采用最广泛的方法之一，常常可以得到很好的baseline性能。



BERT-BILSTM-CRF

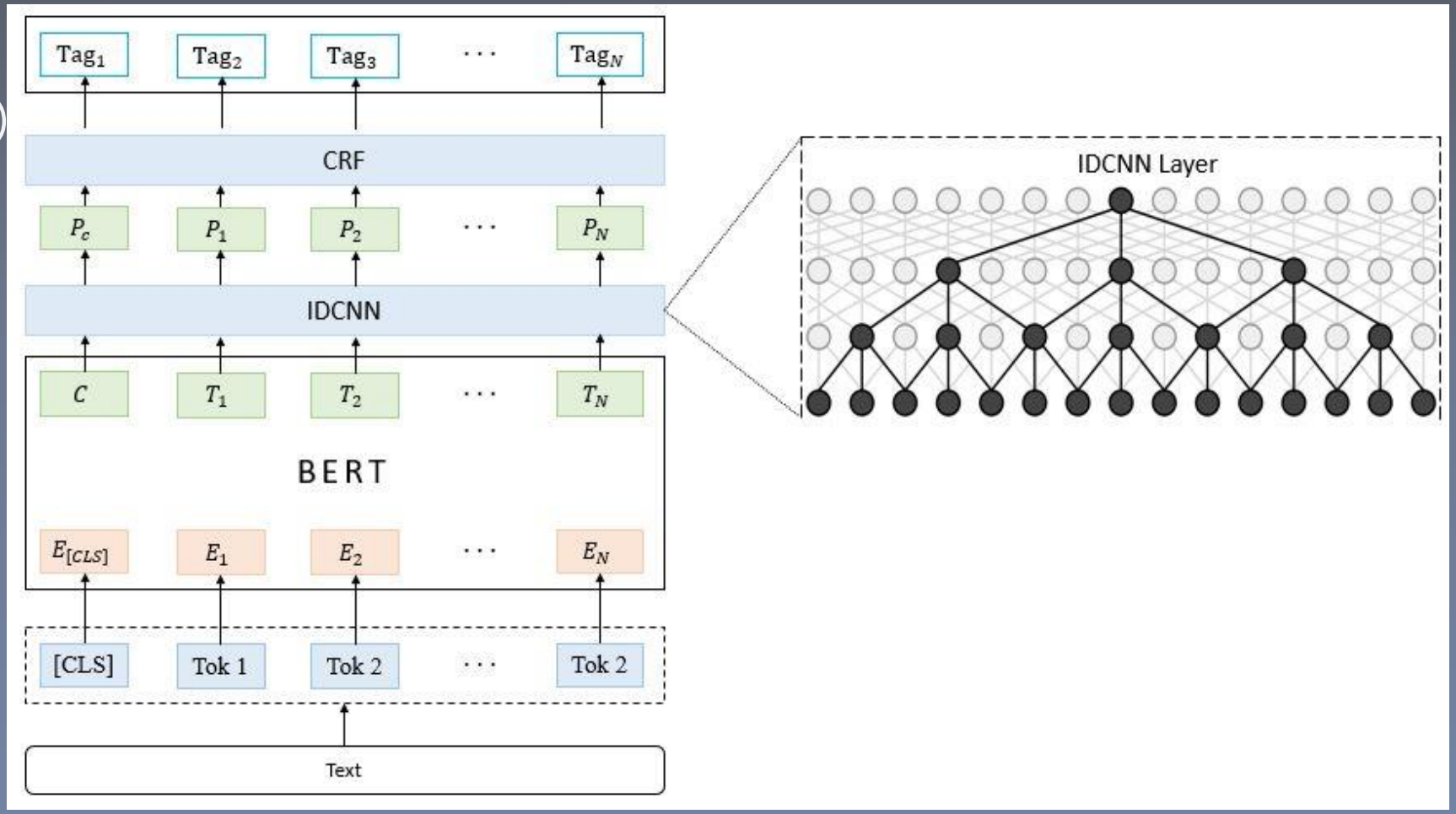
BILSTM-CRF是目前较为流行的命名实体识别模型。可将**BERT预训练模型学习到的token向量输入BILSTM模型进行进一步学习**，让模型更好的理解文本的上下关系，最终通过CRF层获得每个token的分类结果。



BERT-IDCNN-CRF

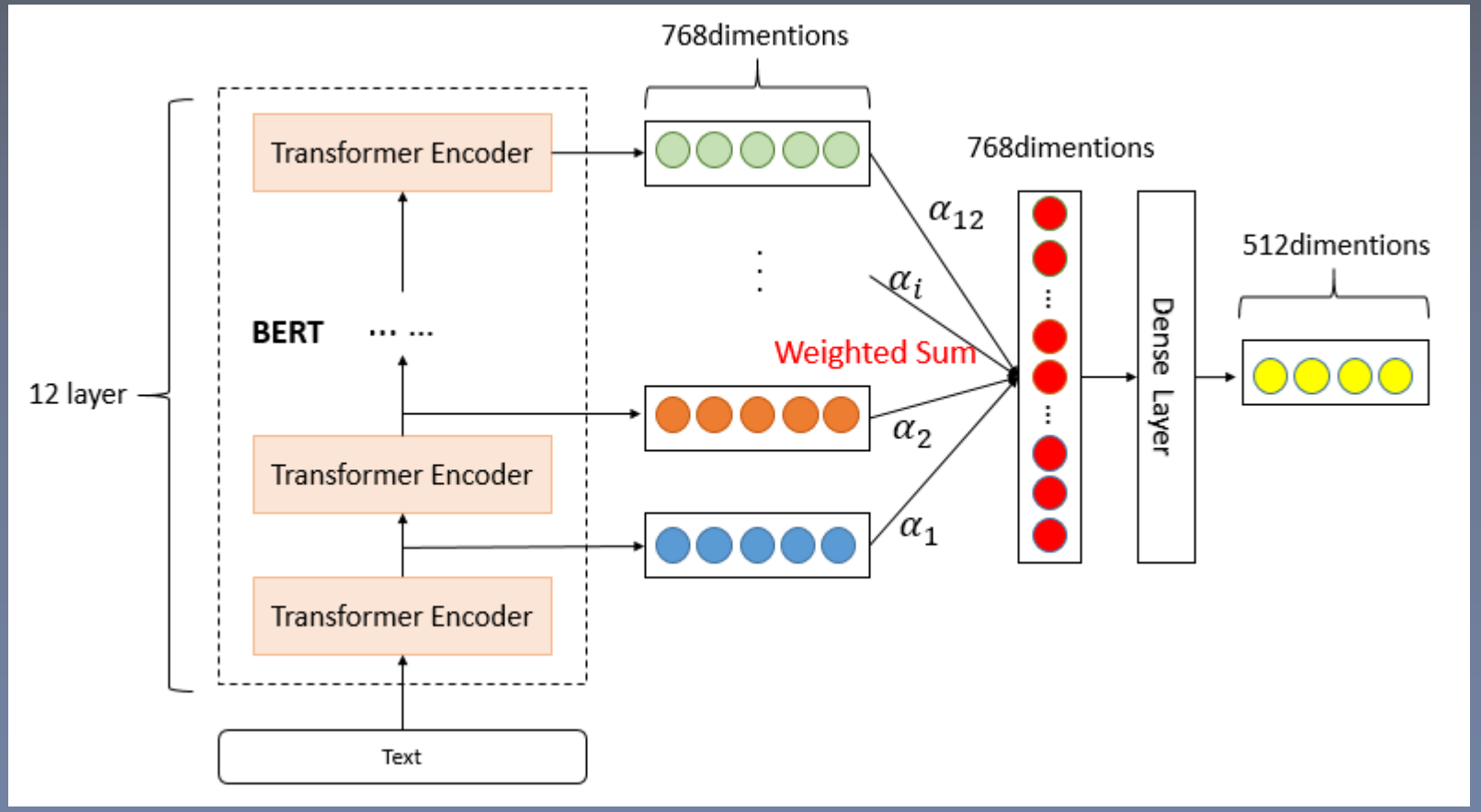
IDCNN通过利用空洞(即补0)来改进CNN结构，在丢失局部信息的情况下，捕获长序列文本的长距离信息，适合当前长文本的数据集。

模型预测速度快，更好的在工业界中使用。



BERT多层表示的动态权重融合

将BERT的十二层transformer生成的表示赋予一个权重,而后通过训练来确定权重值，并将每一层生成的表示加权平均，再通过一层全连接层降维至512维。



bert4keras库串讲

Github: https://github.com/bojone/bert4keras (稳定版本0.8.3)

- basic_extract_features.py: 基础测试，测试BERT对句子的编码序列。
- basic_language_model_gpt2_ml.py: 基础测试，测试GPT2_ML的生成效果。
- basic_language_model_nezha_gen_gpt.py: 基础测试，测试GPT Base（又叫NEZHE-GEN）的生成效果。
- basic_masked_language_model.py: 基础测试，测试BERT的MLM模型效果。
- basic_simple_web_serving_simbert.py: 基础测试，测试自带的WebServing（将模型转化为Web接口）。
- task_conditional_language_model.py: 任务例子，结合 BERT + Conditional Layer Normalization 做条件语言模型。
- task_iflytek_adversarial_training.py: 任务例子，通过对抗训练提升分类效果。
- task_iflytek_bert_of_theseus.py: 任务例子，通过BERT-of-Theseus来进行模型压缩。
- task_iflytek_gradient_penalty.py: 任务例子，通过梯度惩罚提升分类效果，可以视为另一种对抗训练。
- task_image_caption.py: 任务例子，BERT + Conditional Layer Normalization + ImageNet预训练模型 来做图像描述生成。
- task_language_model.py: 任务例子，加载BERT的预训练权重做无条件语言模型，效果上等价于GPT。
- task_reading_comprehension_by_mlm.py: 任务例子，通过MLM模型来做阅读理解问答，属于简单的非自回归文本生成。
- task_reading_comprehension_by_seq2seq.py: 任务例子，通过UniLM式的Seq2Seq模型来做阅读理解问答，属于自回归文本生成。
- task_relation_extraction.py: 任务例子，结合BERT以及自行设计的"半指针-半标注"结构来做关系抽取。
- task_sentence_similarity_lcqmc.py: 任务例子，句子对分类任务。
- task_sentiment_albert.py: 任务例子，情感分类任务，加载ALBERT模型。
- task_sentiment_integrated_gradients.py: 任务例子，通过积分梯度的方式可视化情感分类任务。
- task_sentiment_virtual_adversarial_training.py: 任务例子，通过虚拟对抗训练进行半监督学习，提升小样本下的情感分类性能。
- task_seq2seq_autotitle.py: 任务例子，通过UniLM式的Seq2Seq模型来做新闻标题生成。
- task_seq2seq_autotitle_csl.py: 任务例子，通过UniLM式的Seq2Seq模型来做论文标题生成，包含了评测代码。
- task_sequence_labeling_cws_crf.py: 任务例子，通过 BERT + CRF 来做中文分词。
- task_sequence_labeling_ner_crf.py: 任务例子，通过 BERT + CRF 来做中文NER。

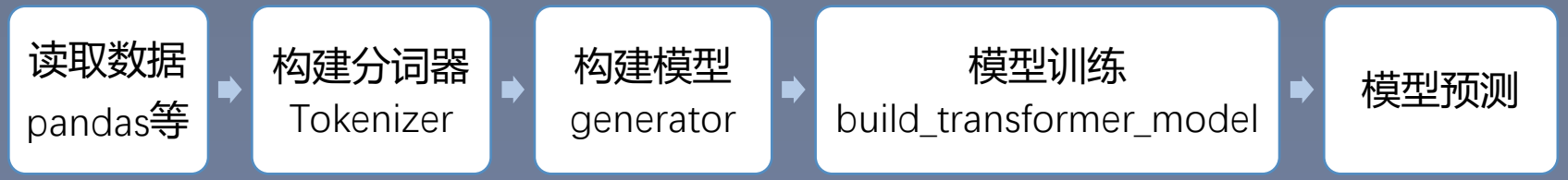
bert4keras库串讲

用法：

目前支持的预训练模型有BERT、ALBERT、roberta、nezha、electra、GPT、T5。

注意：需要了解各种预训练的模型原理以及fine-tune阶段的使用方法。

有时间可以读库中的源码（提升代码水平）



其他类似的库：transformers(https://github.com/huggingface/transformers)

keras-bert(https://github.com/CyberZHG/keras-bert)

深度之眼
deepshare.net

3、解码器串讲

解码器串讲

- Softmax
- **CRF**
- MEMM
- HMM

原理参考：

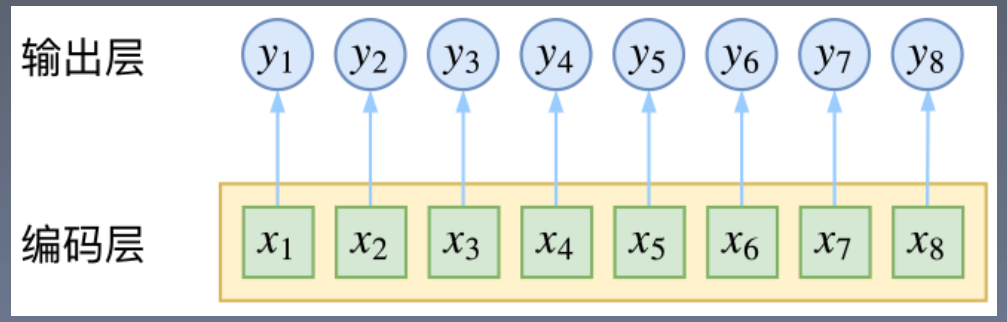
https://kexue.fm/archives/5542

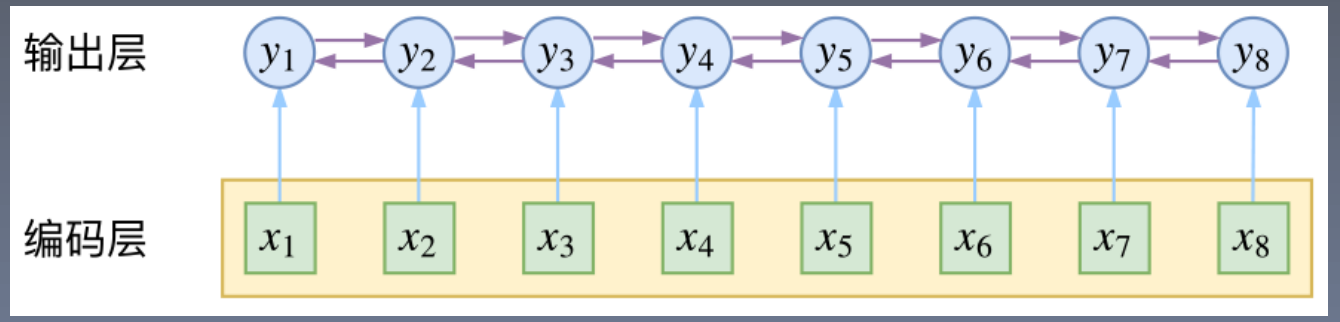
《统计机器学习》

https://kexue.fm/archives/7213

这些原理相对较为复杂！

Softmax vs CRF





CRF的作用： softmax并没有直接考虑输出的上下文关联，而CRF在输出端显式地考虑了上下文关联。

CRF**转移矩阵**示例（代码中讲解）

Softmax vs CRF

当文本语义表征足够好，如Base-BERT、Large-BERT，CRF作用不是很明显，甚至可能不起作用。

可能需要做的实验：

- 去掉CRF层，用softmax解码，对比结果。
- 适当增加CRF层的学习率。

CRF、HMM、MEMM区别

HMM、MEMM、CRF被称为是三大经典概率图模型。

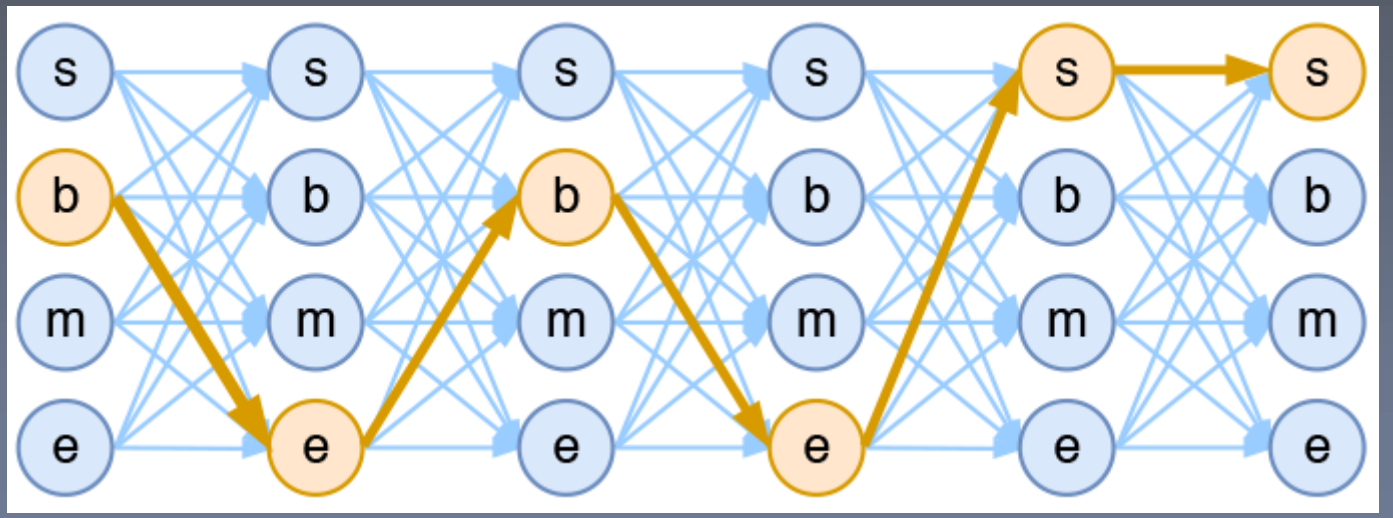
相关原理：https://kexue.fm/archives/7213

一般地，对于NER任务来说，**CRF是最优选择。**

当然也可以进行实验，notebook中演示！

CRF原理浅谈

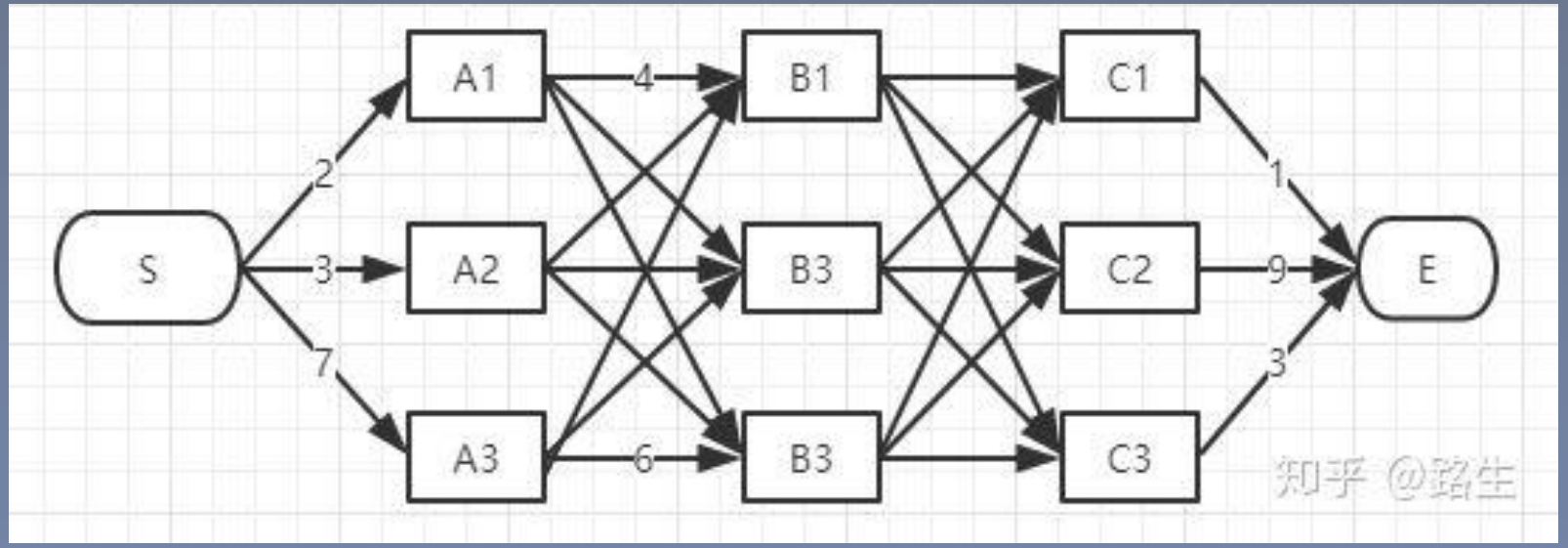
每个点代表一个标签的可能性，点之间的连线表示标签之间的关联，而每一种标注结果，都对应着图上的一条完整的路径。



Viterbi解码

如你从S和E之间找一条最短的路径，除了遍历完所有路径，还有什么更好的方法?

答案：viterbi (维特比)算法。



解码器交叉验证

Bert4keras如何交叉验证?

- 投票
- 概率加权(平均)

解码器交叉验证

投票方法：

对所有单模的文字结果进行投票，可以通过设定阈值，统计每一条数据的预测实体在所有模型的出现次数，当实体出现次数大于阈值时，则认为该实体是预测实体，将其保留。

概率相加：

还原模型输出的概率文本，进行加权。

深度之眼
deepshare.net

比赛相关的知识点

标签缺失、标签错误

知识蒸馏：

训练集有一定缺漏和不规范的，因此，可以尝试一种类似知识蒸馏的方式来重新整理训练集，改善训练集质量。

比如

首先，使用原始训练集加交叉验证的方式，得到了8个模型，然后用这8个模型对训练集进行预测，得到关于训练集的8份预测结果。如果某个样本的某个标签同时出现在8份预测结果中但没有出现在训练集的标注中，那么就将这个标签补充到该样本的标注结果中；如果某个样本的某个标签在8份预测结果中都没有出现但却被训练集标注了，那么将这个标签从该样本的标注结果中去掉。

深度之眼
deepshare.net

面试有哪些需要注意

针对这个比赛，有什么问题

- 各种编码器之间的异同点，如BERT、Roberta、XLNET
- 尝试过的解码器，优缺点对比？
- CRF原理
- 标注质量差、标签缺少问题处理方法
- 如何平衡或提高P、R指标
- 如何解决样本不平衡问题

结 语

感谢大家参加本次的比赛直播

课下请一定要

亲自动手操作一次！

deepshare.net

深度之眼

联系我们:

电话: 18001992849

邮箱: service@deepshare.net

Q Q: 2677693114

公众号

客服微信