NLP 项目相关网址

目 录

第 1 章 问答摘要与推理

1.1 项目课程简介

1. 张楠 github 地址：
https://github.com/HouchangX-AI/Question-and-answer-summary-and-reasoning/blob/master/README.md
2. 《神经网络与深度学习》，其附件中数学基础知识讲解。Github 链接如下：
https://github.com/nndl/nndl.github.io
3. skip-gram 模型博文：https://zhuanlan.zhihu.com/p/27234078
4. 优秀学员 github：https://github.com/Light2077/
5. 推荐一个查阅资料的网站：https://medium.com/ 和 cs224n
6. 华为云 AI 平台：modelarts
7. 百度 AI：黄埔学院
8. 需要补充学些的知识点：交叉熵、反向求导
9. 百度 AI studio：https://aistudio.baidu.com/aistudio/competition/detail/3
10. gitignore 使用：https://github.com/github/gitignore/blob/master/Python.gitignore
11. paperswithcode：
https://github.com/paperswithcode/releasing-research-code/blob/master/templates/README.md
12. 论文 Word2Vec Tutorial - The Skip-Gram Model：
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
13. 论文 Efficient Estimation of Word Representations in Vector Space：
https://arxiv.org/pdf/1301.3781.pdf

1.2 词向量实践

1. 腾讯 800 万中文词的 NLP 数据集开源：https://zhuanlan.zhihu.com/p/47133426
2. 腾讯 AIlib：https://ai.tencent.com/ailab/nlp/embedding.html
3. pycharm 中安装 Conda、pytorch 环境：https://pytorch.org/get-started/locally/
4. 负采样示例：
http://mccormickml.com/2017/01/11/word2vec-tutorial-part-2-negative-sampling/
5. 论文：word2vec Parameter Learning Explained.pdf

1.3 Seq2Seq 一

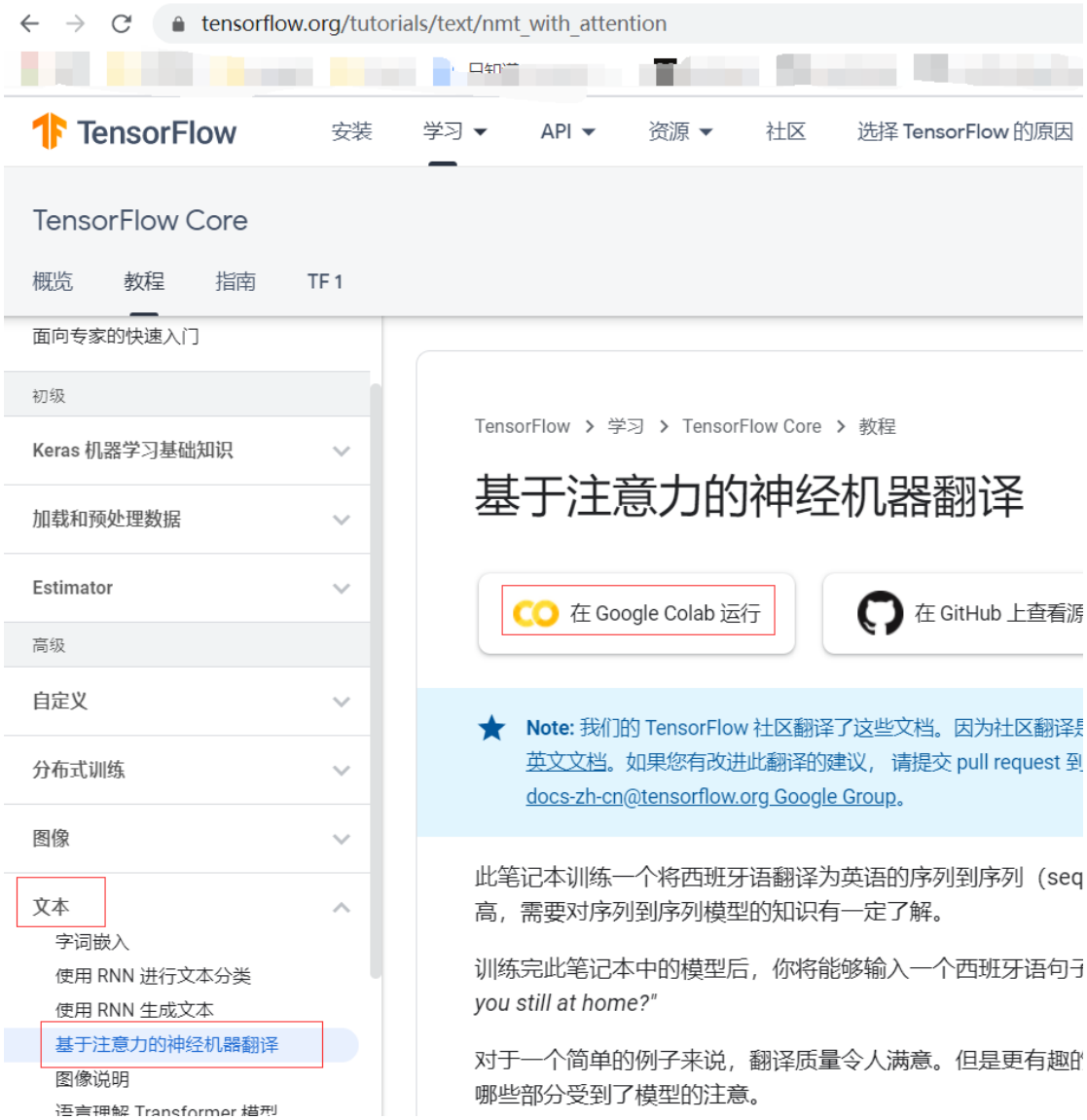
1. OpenNMT：Open source ecosystem for neural machine translation and neural sequence learning：https://github.com/OpenNMT

2. Encoder、Decoder、Attention 层解说博文：
   https://blog.csdn.net/zimiao552147572/article/details/105893842

3. TensorFlow 官方 attention 文档“基于注意力的神经机器翻译”：
   https://www.tensorflow.org/tutorials/text/nmt_with_attention

4. Google Cloab：
   https://colab.research.google.com/github/tensorflow/docs-l10n/blob/master/site/zh-cn/tutorials/text/nmt_with_attention.ipynb

注：从如下接口进入 Google Cloab



5. Sequence Modeling: Recurrentand Recursive Nets：

http://www.deeplearningbook.org/contents/rnn.html

6. Attention and Augmented Recurrent Neural Networks：
https://distill.pub/2016/augmented-rnns/

1.4 Seq2Seq 二

1. 后厂理工 ai github：https://github.com/HouchangX-AI
2. 微分代码 github：
https://github.com/ZhaoYi1031/automatic_differentiation/blob/master/autodiff_test.py
3. 后厂理工代码参考：
https://github.com/HouchangX-AI/Question-and-answer-summary-and-reasoning
4. 优化器：https://deeplearning.ai/ai-notes/optimization/
5. 网络参数初始化：https://www.deeplearning.ai/ai-notes/initialization/
6. 华为云：https://auth.huaweicloud.com/authui/login.html#/login
7. python argparse 讲解：https://zhuanlan.zhihu.com/p/28871131
8. 论文：Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks
9. 论文：DIVERSE BEAM SEARCH.pdf
10. 论文：ROUGE- A Package for Automatic Evaluation of Summaries.pdf
11. 论文:Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks.pdf

1.5 项目模型算法提升

1. transformer 模型 github：https://github.com/huggingface/transformers
2. 文章：BERT-Pre-training of Deep Bidirectional Transformers for Language Understanding：https://arxiv.org/pdf/1810.04805.pdf
（翻译博文：https://blog.csdn.net/sinat_33741547/article/details/86311310）
3. bert 模型 github：https://github.com/jihun-hong/Bert-Classifier/tree/master/src/models
4. 论文：Incorporating Copying Mechanism in Sequence-to-Sequence Learning
5. 论文：Get To The Point: Summarization with Pointer-Generator Networks
6. 论文：Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks
7. 论文：Convolutional Sequence to Sequence Learning
8. 论文：Transformers and Pointer-Generator Networks for Abstractive Summarization
9. 论文：Text Summarization Techniques: A Brief Survey
10. 论文：Fine-tune BERT for Extractive Summarization

11. 论文：Get To The Point- Summarization with Pointer-Generator Networks.pdf

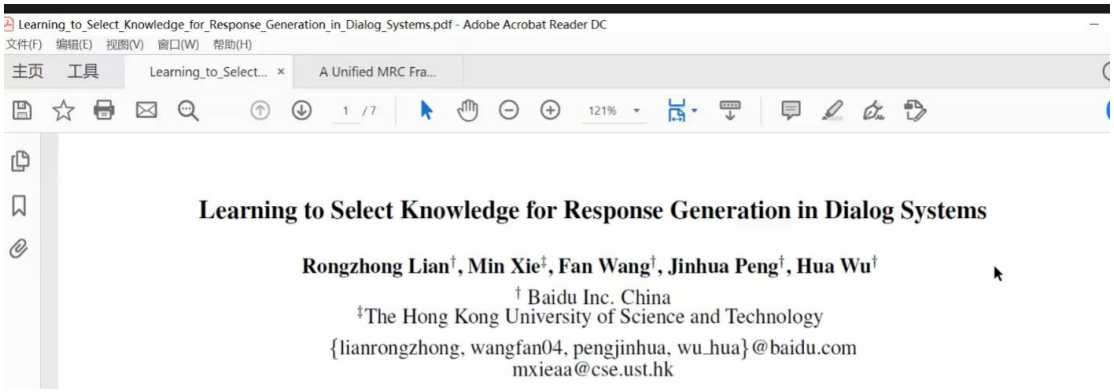
1.6 项目代码部署与总结

1. 中文预训练模型牛人 github：https://github.com/bojone/bert4keras
2. 博客：http://jalammar.github.io/illustrated-transformer/
3. nlp 顶会论文：THE CURIOUS CASE OF NEURAL TEXT DeGENERATION
4. 部署模型：tensorflow.org/tfx/tutorials/serving/rest_simple
5. 牛人（brightmaret）github：https://github.com/bojone/albert_zh
   https://github.com/brightmart/albert_zh
6. transformer 作者 github：
   https://github.com/huggingface/transformers/blob/master/examples/summarization/bertabs/modeling_bertabs.py
   https://github.com/huggingface/transformers
7. docker：https://hub.docker.com/
8. GPT2: https://github.com/qingkongzhiqian/GPT2-Summary
9. opennmt：https://opennmt.net/
10. 论文：the_curious_case_of_neural_text_degeneration.pdf
11. cnn-dailymail 数据集地址：https://github.com/abisee/cnn-dailymail
12. 论文：Transformers and Pointer-Generator Networks for Abstractive Summarization
13. 论文：Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks
14. 论文：Fine-tune BERT for Extractive Summarization

第 2 章 试题知识点多标签分类

2.1 项目简介

1. 试题数据：http://tiku.21cnjy.com/tiku.php?mod=quest&channel=8&cid=1155&xd=2
2. 百度 aistudio：https://aistudio.baidu.com/aistudio/projectdetail/241741
3. 论文：Learning to Select Knowledge for Response Generation in Dialog Systems



4. 工程数据集：链接：https://pan.baidu.com/s/1RYJYbhSP9wpbutFUSAHg8Q
   提取码：1ofn
5. 基于机器阅读理解框架的命名实体识别方法：A Unified MRC Framework for Named Entity Recognition （ACL2020）

NER：Named Entity Recognition 命名实体识别

2.2 常用机器学习

1. 看论文网站：

   http://arxiv-sanity.com/

   https://paperswithcode.com/

   webofscience

   ACL 会议

2. bert 牛人 github：https://github.com/bojone/bert4keras
3. keras api：https://keras.io/zh/layers/core/
4. bert 博文（川大的）：https://blog.csdn.net/weixin_42001089/article/details/97657149
5. bert4keras 博文：https://kexue.fm/archives/7161
6. 多分类评价指标 知乎文章：https://zhuanlan.zhihu.com/p/64315175
7. 正则化通常加在计算损失函数中。正则化方法介绍：

   https://baike.baidu.com/item/正则化方法/19145625?fr=aladdin
8. 论文：Global Vectors for Word Representation
9. 论文：Convolutional Neural Networks for Sentence Classification
10. 论文：A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification
11. TextCNN 论文

- https://arxiv.org/pdf/1408.5882.pdf
- https://arxiv.org/pdf/1510.03820.pdf

12. GLOVE & FastText 论文：

- https://nlp.stanford.edu/pubs/glove.pdf
- https://arxiv.org/pdf/1607.01759.pdf

2.3 Transformer ELMO GPT

1. self-attention：
https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a#570c

2.transformer attention：
https://medium.com/@bgg/seq2seq-pay-attention-to-self-attention-part-2-中文版-ef2ddf8597a4

3. ELMO 原理解析：https://zhuanlan.zhihu.com/p/51679783

4. tf.keras.layers.LayerNormalization：
https://www.tensorflow.org/api_docs/python/tf/keras/layers/LayerNormalization

5. rasa community：https://rasa.com/showcase/dialogue-virtual-clinic

6. chatito 工具：https://github.com/rodrigopivi/Chatito

7. Neo4j：http://49.233.155.170:7474/browser/

注：Neo4j 需要自己构建，参考 github：
https://github.com/pengyou200902/Doctor-Friende

8. rasa 中文聊天机器人开发指南：
https://blog.csdn.net/AndrExpert/article/details/104328946

9. rasa 官网：https://rasa.com/showcase/

10. 残差网络：https://zhuanlan.zhihu.com/p/80226180

11. 论文：2017 Google : Attention is all you need

12. 论文：Recurrent Models of Visual Attention

13. 论文：Neural Machine Translation by Jointly Learning to Align and Translate

14. 论文：Neural Machine Translation by Jointly Learning to Align and Translate

15. 论文：Deep contextualized word representations（NAACL）

16. 论文：ELMo：Embedding from Language Model

17. 论文：OpenAI 2018 ：Improving Language Understanding by Generative Pre-Training （GPT 1）

18. Transformer 论文：https://arxiv.org/pdf/1706.03762.pdf

19. ELMo 论文：https://arxiv.org/pdf/1802.05365.pdf

20. GPT 论文：

https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

2.4 bert 结构详解

1. attention 综述：https://zhuanlan.zhihu.com/p/62136754
2. bert 简介：https://zhuanlan.zhihu.com/p/92849070
3. 自认语言处理任务评价网站：https://gluebenchmark.com/leaderboard
4. bert 源码 github：https://github.com/google-research/bert
5. bert-as-service 包：https://github.com/hanxiao/bert-as-service
6. bert 论文：https://arxiv.org/pdf/1810.04805.pdf
7. bert 解读博客：https://blog.csdn.net/weixin_42001089/article/details/97657149
8. ai 竞赛开放平台：https://www.flyai.com/
9. 图神经网络工具：https://github.com/thunlp/OpenKE
10. 论文：BERT：Pre-traning of Deep Bidirectional Transformers for Language Understanding（Google AI Language）
11. 论文：BERT： Bidirectional Encoder Representations from Transformer
12. GLUE 模型排名：https://gluebenchmark.com/leaderboard
13. 数据集：BooksCorpus and Wikipedia 数据集
14. 论文：ACL 2019：What does BERT learn about the structure of language?：https://hal.inria.fr/hal-02131630/document
15. bert 源码 github：https://github.com/google-research/bert
16. bert 源码 github：https://github.com/google-research/bert
17. 论文：BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding：https://arxiv.org/pdf/1810.04805.pdf
18. 论文：BERT Pre-training of Deep Bidirectional Transformers for Language Understanding.pdf

2.5 ERNIE 与 XLNet 介绍

1. OpenKE 网址：http://139.129.163.161//
2. TransE（translating embeddings）：https://zhuanlan.zhihu.com/p/32993044
3. OpenKE github：https://github.com/thunlp/OpenKE/tree/OpenKE-PyTorch/openke/module/model
4. GCN 图卷积神经网络：https://zhuanlan.zhihu.com/p/72546603
5. GCN 在文办分类中的应用：https://juejin.im/post/5d90658ae51d4578331cbce5

6. scipy csr_matrix 和 csc_matrix 函数详解：
https://blog.csdn.net/u013010889/article/details/53305595
7. 论文：Hierarchical Taxonomy-Aware and Attentional Graph Capsule RCNNs for Large-Scale Multi-Label Text Classification
8. 论文：Graph Convolutional Networks for Text Classification
9. paddle github：https://github.com/PaddlePaddle/PaddleHub
10. 论文：ERNIE: Enhanced Representation through Knowledge Integration （Baidu）
11. 论文：ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding （Baidu）
12. Paddle 相关网址：

- https://github.com/PaddlePaddle/Paddle
- https://www.paddlepaddle.org.cn/
- https://nlp.baidu.com/homepage/nlptools/

13. paddle github：
https://github.com/PaddlePaddle/PaddleHub/tree/release/v1.5/demo/text_classification
14. 利用 paddle 进行文本情感分类：
https://nlp.baidu.com/homepage/nlptools/document?f=文本情感分类&sd=1576344688395
15. ERNIE github：https://github.com/PaddlePaddle/ERNIE
16. ERNIE 1.0 论文：https://arxiv.org/pdf/1904.09223.pdf
17. ERNIE 2.0 论文：https://arxiv.org/pdf/1907.12412v1.pdf
18. 论文：Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context（Google）
19. XLNet 论文：XLNet：Generalized Autoregressive Pretraining for Language Understanding（Google）
20. Transformer-XL 论文：https://arxiv.org/pdf/1901.02860.pdf
21. XLNet 论文：https://arxiv.org/pdf/1906.08237.pdf
22. 中文 XLNet 预训练模型：https://github.com/ymcui/Chinese-PreTrained-XLNet
23. 论文：ERNIE 2.0 A CONTINUAL PRE-TRAINING FRAMEWORK FOR LANGUAGE UNDERSTANDING.pdf
24. 论文：vERNIE Enhanced Representation through Knowledge Integration.pdf

2.6 模型部署与课程总结

1. 算法及代码跟进网站：https://paperswithcode.com/
2. docker：https://hub.docker.com/
3. paperswithcode ：https://paperswithcode.com/
4. bert4keras 模型：https://kexue.fm/archives/7161
5. bert 实践：关系抽取解读博客：
   https://blog.csdn.net/weixin_42001089/article/details/97657149
6. 数据竞赛社区：https://biendata.xyz/
7. 论文：StructBert: Incorporating Language Structures Into Pre-Training For Deep Language Understanding（Alibaba 2019）
8. 论文：Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer（Google 2019）
9. Struct-Bert 论文：https://arxiv.org/pdf/1908.04577.pdf
10. T5 参考资料：
    - https://arxiv.org/abs/1910.10683
    - https://github.com/google-research/text-to-text-transfer-transformer
    - https://zhuanlan.zhihu.com/p/89719631
    - https://zhuanlan.zhihu.com/p/88727133
    - https://zhuanlan.zhihu.com/p/88438851
11. Tf-serving：
    - https://github.com/tensorflow/serving
    - https://www.tensorflow.org/tfx/guide/serving

第 3 章 试题知识点多标签分类

3.1 基础框架和评价指标

1. DuReader 数据集：https://ai.baidu.com/broad/download
2. 中文预训练模型：https://linux.ctolib.com/brightmart-albert_zh.html
3. bert4keras albert：
   https://github.com/bojone/bert4keras/blob/master/examples/task_sentiment_albert.py
4. 论文：Rethinking Batch Normalization in Transformers
5. 论文：Understanding and Improving Layer Normalization

6. 参考文献：

- He W, Liu K, Liu J, et al. Dureader: a chinese machine reading comprehension dataset from real-world applications[J]. arXiv preprint arXiv:1711.05073, 2017.
- Liu S, Zhang X, Zhang S, et al. Neural machine reading comprehension: Methods and trends[J]. Applied Sciences, 2019, 9(18): 3698.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
- Shen S, Yao Z, Gholami A, et al. Rethinking Batch Normalization in Transformers[J]. arXiv preprint arXiv:2003.07845, 2020.
- Xu J, Sun X, Zhang Z, et al. Understanding and Improving Layer Normalization[C]//Advances in Neural Information Processing Systems. 2019: 4383-4393.

3.2 序列标注与 Attention 机制

1. pytorch-crf 库：https://pytorch-crf.readthedocs.io/en/stable/
2. 论文 Log-Linear Models, MEMMs, and CRFs：
   http://www.cs.columbia.edu/~mcollins/crf.pdf

3.3 双向 Attention

1. 论文：Hierarchical attention flow for multiple-choice reading comprehension
2. 中文文本纠错 github：https://github.com/shibing624/pycorrector
3. bert4keras：https://github.com/bojone/bert4keras/tree/master/examples
2. 代码下载链接：https://pan.baidu.com/s/1Tm4xeqTnbSKcXHd92Qo4Gg
   提取码：5mpm

3.4 Multi-Hop 机制

1. 论文：Iterative alternating neural attention for machine reading
2. 论文：Machine comprehension using match-lstm and answer pointer
3. 论文：Gated-Attention Readers for Text Comprehension
4. 论文：Gated Self-Matching Networks for Reading Comprehension and Question Answering
5. 论文：R-NET: Machine reading comprehension with self-matching networks

6. 论文：Dcn+: Mixed objective and deep residual coattention for question answering
7. 论文：S-net: From answer extraction to answer synthesis for machine reading comprehension

3.5 Multi-Hop 机制与 Memory Network

1. 论文：Multi-hop Reading Comprehension across Documents with Path-based Graph Convolutional Network：https://arxiv.org/pdf/2006.06478.pdf

3.6 Decoder and MRC Trick

1. 2020 语言与智能技术竞赛：机器阅读理解任务：
http://colab.research.google.com/drive/1P3MPZJ7iKYc8hGe5IPY7CUj9gBMphLO1#scrollTo=aBwxu8Lw83Up
http://colab.research.google.com/drive/1P3MPZJ7iKYc8hGe5IPY7CUj9gBMphLO1#scrollTo=_IIY1GnV83T3
2. colab：
http://colab.research.google.com/drive/1P3MPZJ7iKYc8hGe5IPY7CUj9gBMphLO1#scrollTo=b4aRVkx883T4
3. google 网盘：http://drive.google.com/drive/my-drive
4. 张寓弛知乎：https://www.zhihu.com/people/johnny-richards/answers
5. 张寓弛知乎 https://www.zhihu.com/question/54504471/answer/630639025

第 4 章 数据集网址

1. 数据集：BooksCorpus and Wikipedia 数据集
2. DuReader 数据集：https://ai.baidu.com/broad/download
3. cnn-dailymail 数据集地址：https://github.com/abisee/cnn-dailymail
4. 爱奇艺文本纠错数据集
5. 搜狗文本纠错数据集
6. 人民日报数据集：http://s3.bmio.net/kashgari/china-people-daily-ner-corpus.tar.gz
7. CCKS2018 比赛，数据包含训练和复赛测试数据以及测试脚本：
   链接：https://pan.baidu.com/s/16Ue4ZvRJuE9TWdsKWYuxbg
   提取码：7kwp
8. https://github.com/shibing624/pycorrector#数据集下载

- 人民日报中文语料库
- NLPCC 2018 GEC 官方数据集
- NLPCC 2018+HSK 熟语料
- NLPCC 2018+HSK 原始语料

9. 中国古诗词数据集
10. 项目 NER 数据集：https://trello.com/c/jp19knxc/25-ner 数据地址