NLP 入门指导

Lanco Lab

February 20, 2019

1 Lanco 实验室实习要求

1.1 实习要求

- 按时完成老师布置的相关的工作。
- 大约**每 2 周**做一次正式的实习进度报告，几十字即可。
- 平时多向指导自己的师兄师姐汇报进度（如一周一次），讨论解决遇到的问题。

1.2 入门建议

- 优先完成下面介绍的**入门项目**，做到能够**快速搭建自己的网络架构**。
- 在保证入门项目的进度的情况下，可以夯实自己的基础。可以学习深度学习知识优先（推荐吴恩达的 deepai 的系列课程）；然后是吴恩达关于机器学习的新版课程；深度学习、强化学习和机器学习的其他学习资料作为参考。
- 编程语言希望熟练掌握**python**，深度学习框架实验室成员多使用**pytorch**。

1.3 入门资料

下面详细介绍了 NLP 的入门项目以及基础知识两部分，以**优先完成入门项目**为重。

2 入门项目

入门项目主要包括文本分类和文本生成两部分，此外还有一些可选内容。

2.1 文本分类

2.1.1 任务及数据集

- **数据集：** Amazon Dataset（5 分类）
- **下载链接：**https://drive.google.com/file/d/125WOYx6G18yk_uFn52jExlFv9hIM99xx/view?usp=sharing

2.1.2 模型实现

1. **机器学习算法：** Logistic, SVM, Boosting 等各种方法; **可以自己切分小部分 Amazon 数据集做实验**
   - 可以调 sklearn 包，特征可以使用 tf-idf 或者 word2vec 向量
2. **深度学习模型：** CNN, LSTM, C-LSTM；**在 Amazon 全数据集上做实验**
   - 上述三种模型必须实现，同时实现在上述三种模型中分别加入 self-attention 机制的版本。
   - **可选：**尽力实现 Hierarchical Attention 用于文本分类。

2.1.3 参考论文

- 《Convolutional Neural Networks for Sentence Classification》
- 《Character-level Convolutional Networks for Text Classification》
- 《A C-LSTM Neural Network for Text Classification》
- 《Hierarchical Attention Networks for Document Classification》

2.2 文本生成

2.2.1 任务及数据集

- **机器翻译：**英语-越南语
- **下载链接：**https://drive.google.com/file/d/1L7j_gqF1dN49BD3l4km95p0nN5ilwNAR/view?usp=sharing

2.2.2 模型实现

1. **Seq2Seq 模型：**
   - 使用不同的 attention；使用 greedy search 以及 beam search；尝试使用 BPE(**可选**)
   - 阅读组内相关文本生成的论文，每篇论文基本都有相关的开源代码。
   - **实验室 github:** https://github.com/lancopku
2. **Transformer 模型：**
   - **可选：**尽力实现，可以借助 OpenNMT 的框架。但是基本原理一定要明白。
   - **OpenNMT 网址：**http://opennmt.net/

2.2.3 参考论文

- 《Neural Machine Translation by Jointly Learning to Align and Translate》
- 《Sequence to Sequence Learning with Neural Networks》
- 《Effective Approaches to Attention-based Neural Machine Translation》
- 《Asynchronous Bidirectional Decoding for Neural Machine Translation》
- 《Convolutional Sequence to Sequence Learning》
- 《Attention Is All You Need》
- 实验室已发的关于 Seq2Seq 模型的一些论文

2.3 GAN 与强化学习

可以不实现相应的框架，但一定要清楚相关原理。在完成上面两个必备任务的基础上，可以阅读如下参考论文。

- 《Generative Adversarial Networks》
- 《SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient》
- 《SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks》
- 《DP-GAN: Diversity-Promoting Generative Adversarial Network for Generating Informative and Diversified Text》(本组许晶晶师姐论文)
- 《Improving Neural Machine Translation with Conditional Sequence Generative Adversarial Nets》
- 《Sequence Level Training with Recurrent Neural Networks》
- 《Adversarial Learning for Neural Dialogue Generation》
- 《Deep reinforcement learning for dialogue generation》
- 《Self-Critical Sequence Training for Image Captioning》
- 《An Actor-Critic Algorithm for Sequence Prediction》
- 《A Deep Reinforced Model for Abstractive Summarization》

3 基础知识

3.1 Python 学习资料

- **小甲鱼零基础入门 python：**教学视频，共 96 节课，看前 63 节即可。
- **链接：**https://pan.baidu.com/s/1ihxScS4xNUx84WMxOMaQeA **密码：**8mbe

3.2 Pytorch 学习资料

- **网络博客：**https://morvanzhou.github.io/tutorials/machine-learning/torch/
- **英文官网：**https://pytorch-cn.readthedocs.io/zh/latest/
- **中文官网：**https://pytorch.org/

3.3 深度学习资料

3.3.1 教学视频

- **吴恩达（deepai 五门课程）：**https://mooc.study.163.com/smartSpec/detail/1001319001.htm
- **李飞飞（CS231n）：**http://study.163.com/course/introduction.htm?courseId=1003223001
- **NLP 课程（CS224n）：**https://www.bilibili.com/video/av15892671/

3.3.2 相关书籍

- **《Deep Learning》英文版：**https://github.com/janishar/mit-deep-learning-book-pdf
- **《Deep Learning》中文版：**https://github.com/exacity/deeplearningbook-chinese

3.3.3 入门博客

- **链接：**https://www.zybuluo.com/hanbingtao/note/433855

3.4 强化学习资料

3.4.1 教学视频：

- **David Sliver：** https://search.bilibili.com/allkeyword=david%20silver&from_source=banner_search

3.4.2 相关书籍：

- Sutton 的《Reinforcement Learning: An Introduction》
- **链接：**http://incompleteideas.net/book/the-book-2nd.html
- David Sliver 的强化学习课程的 PPT

3.5 机器学习资料

3.5.1 教学视频

机器学习比较好的教学视频有吴恩达的老版（斯坦福 CS229 课程）和新版（Coursera 的 Machine Learning 课程），可以先看新版，再看老版。新版课程比较容易，公式推导并不多，偏重应用；老版课程理论证明更多一些。

3.5.2 老版课程

课程配有中文字幕，看起来比较方便。

视频链接：http://open.163.com/special/opencourse/machinelearning.html

3.5.3 新版课程

Coursera 上的课程配有相关的作业，需要使用 Matlab 来实现相应的程序；如果 Coursera 没法看的话可以看哔哩哔哩上的视频，两个 link 里的视频都是一样的。相关视频都有中文字幕。

Coursera：https://www.coursera.org/learn/machine-learning

哔哩哔哩：https://www.bilibili.com/video/av9912938?from=search&seid=11774433852524662823

3.6 相关书籍

《数学之美》是类似于科普性的书籍。《统计学习方法》和《机器学习》是两本比较好的国内的关于机器学习的书，后者内容偏多，类似于综述性文献，可以先看《统计学习方法》再看《机器学习》。《The Elements of Statistical Learning》感觉是机器学习里最经典的著作，各种问题都讲的很清楚，但是内容很多，如果有时间的话可以看一遍。后面的三本书可以当成参考资料。

- 吴军，《数学之美》
- 李航，《统计学习方法》
- 周志华，《机器学习》
- Trevor Hastie，《The Elements of Statistical Learning》，附件有英文版和中文版。
- Keven P. Murphy 《Machine Learning: A Probabilistic Perspective》，附件有英文版。
- Christopher M. Bishop 《Pattern Recognition and Machine Learning》，附件有英文版和中文版。
- Shai Shalev-Shwartz，《Understanding Machine Learning: From Theory to Algorithms》，附件有英文版。

注：机器学习的内容可以留着慢慢看，不是当下最重要的。可以用于在以后的学习中慢慢打基础。