Standford CS336: Language Modeling from Scratch

CS336 学习笔记

中科院信工所 · 杨桂淼

2026 · 02 · 11

千磨万击还坚劲，任尔东西南北风

——献给自己

目录

Part I

Introduction

第 1 章

预备知识与工具

一方面直觉非常重要, 可是另一方面又要能及时吸取新的观念修正自己的直觉.

— 杨振宁 ·《我的学习与研究经历》

Part II

Tokenization

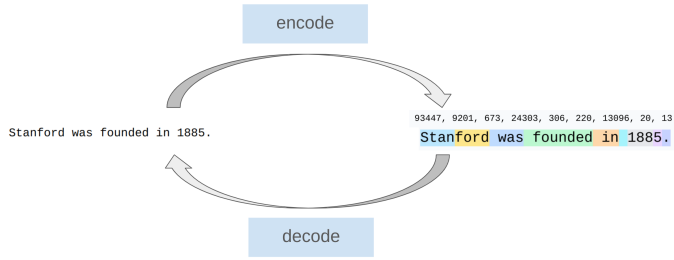

图 1.1: Tokenization 过程

Tokenizers 是将字符串和整数序列之间进行相互转换，四类主要的分词器：character_tokenizer、byte_tokenizer、word_tokenizer、bpe_tokenizer。

character_tokenizer：每个字符可以通过 ord 函数完成字符到 ASCII 码 (Unicode) 的相互转换。但是带来的问题就是构成的词表非常大，许多字符生僻少用导致词汇使用效率低。

byte_tokenizer：Unicode 字符可以被表示为一个使用 0-255 整数表示的字节序列。但是由于一个字节只能被 256 个整数所表示，这就造成了经过 Tokenization 后的序列长度会变得非常长。

word_tokenizer：NLP 领域内常用的方法是将字符串分割为单词，利用正则表达式构建分词器，将基于分词器得到的 segments 映射为整数，构建出每个 segment 的整数映射。但是存在的问题是单词的数量巨大，无法提供固定的词汇量；训练期间未使用过的新词会被标记为 UNK，扰乱困惑度计算。

联想语言模型的输出：LLM 的输出本质上就是在一个大的词表上预测单词的概率分布，如果词表非常大，那么计算了可想而知是非常恐怖的，而且也会影响 token 的预测。上面介绍的三种 Tokenization 都各有特点，但是也都存在致命的缺点。LLM 基于上述编码的劣势，最终选择采用 BPE 算法作为模型的 Tokenization 方法。

第 2 章

Byte Pair Encoding

BPE 算法是在 1994 年发表的“A New Algorithm for Data Compression”提出的一种新型数据压缩算法。核心思想就是将序列中常见的一对相邻的数据单元替换为数据中没有出现过的新单元，反复迭代直至满足终止条件。

bpe_tokenizer：基本思路是常见的字符序列被单个 token 表示，罕见的序列被多个 tokens 表示。

BPE 的基本计算路程如下所示：

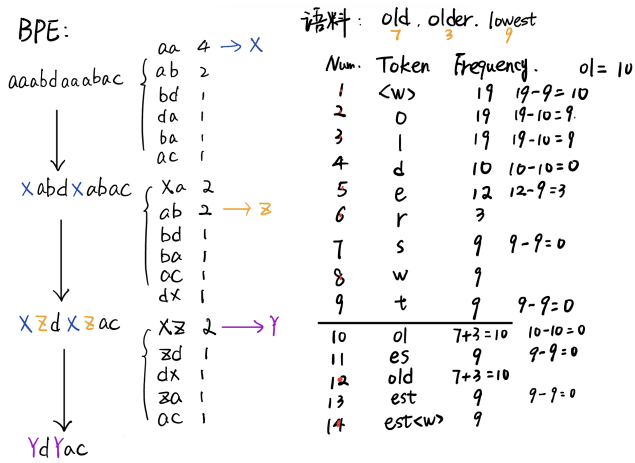
BPE:
aaabdaaabac
XabdXabac
XZdXZac
YdYac
aa 4 → X
ab 2
bd 1
da 1
ba 1
ac 1
Xa 2
ab 2 → Z
bd 1
ba 1
aC 1
dX 1
XZ 2 → Y
Zd 1
dX 1
Za 1
aC 1
语料: old, older, lowest
7 3 9
Num. Token Frequency. ol= 10
1 <w> 19 19-9=10
2 o 19 19-10=9
3 l 19 19-10=9
4 d 10 10-10=0
5 e 12 12-9=3
6 r 3
7 s 9 9-9=0
8 w 9
9 t 9 9-9=0
10 ol 7+3=10 10-10=0
11 es 9 9-9=0
12 old 7+3=10
13 est 9 9-9=0
14 est<w> 9

图 2.1: BPE 算法流程

第 3 章

Homework1:BPE Tokenizer

源码之前，了无秘密；天下大事，必做于细.

— 侯捷 ·《STL 源码剖析》

Part III

Systems

第 4 章

预备知识与工具

想要得到一件东西，最好的办法是让自己配得上它.

— Charles Munger ·《穷查理宝典》

Part IV

Scaling

第 5 章

预备知识与工具

下定决心，排除万难，去争取胜利！

— 毛泽东 · 《愚公移山》

Part V

Data

第 6 章

预备知识与工具

坚持相信的，相信坚持的.

— 王坚 ·《在线》

Part VI

结束语

第 7 章

江湖再见

> 未来可能讲更有意思的话，著更其完美的文，做更加壮丽的事业，但今天只是今天，未来也只是今天的未来。
>
> — 习近平 ·《摆脱贫困》

行文至此，我终于完成了 CS336 全部章节的学习梳理与课后作业。一路坚持，实属不易。若你作为读者能够读到这里，我由衷感谢你的陪伴，相信你也收获颇丰。

我本人目前是中国科学院大学网络空间安全学院的研二学生，培养单位在中国科学院信息工程研究所。自本科阶段起，我便对数据科学与人工智能产生了浓厚兴趣，学习于我而言，始终是兴趣驱动的过程。能够从河北大学保送至国科大继续深造，我深感幸运，也倍加珍惜！

在国科大读书期间接触了很多优秀的大牛老师和优秀的同辈们，他们思考问题的方式与优秀的学习品质都非常值得我学习。

犹记在雁栖湖校区上课时，杨力祥老师的《C++程序设计》令我如沐春风。本科时我总以为，只要足够努力、肯下功夫，就没有学不会的知识；但来到国科大，听了杨老师的课，我的思维方式悄然发生了转变。大师的点拨，往往四两拨千斤，对我后续的学习与发展产生了深远影响。杨老师曾带领团队自主设计操作系统，其著作《Linux 内核设计的艺术》至今仍是我时常翻阅的案头书，书中每一张图解都细致入微，可见其用心之深。从杨老师身上，我学到的不仅是知识，更是如何思考问题、如何解决问题的路径。而这，也正是我整理这份笔记的初衷。（如果你也是国科大的学子，强烈推荐去听一听杨老师的课！）

过去我总认为，只要有兴趣驱动，再投入足够的时间和精力，就一定

能学有所成。然而，进入国科大之后，我愈发意识到：交流与实践，才是最高效的学习方式，尤其是交流。在实践中尝试，在交流中碰撞，在比较中反思，在反复中精进，在曲折中实现螺旋式上升。

知识源于日常的点滴积累，能力需要在工作中耐心打磨，技术得益于实践中的总结提升。在如今这个流媒体盛行的时代，能够静下心来踏踏实实钻研技术，已愈发难得。当所有人都在追求“更快”的时候，我反而想试着“慢”下来，哪怕只是慢一点。因为比起一味求快，我更害怕因匆忙而频频“摔跤”。或许我做不了时代的弄潮儿，但我愿始终做一名坚定的追随者。

一生很长也很短，能做点有意思、有价值的事儿，我觉着很 TMD 值得！**“若留下探索，后人总结；若留下经验，后人咀嚼；若留下教训，后人借鉴；若留下失误，后人避免。我亦断定此书会被人遗忘。遗忘乃是大好事，足以证明我们前进得很快。”**

最后，感谢提供 LaTeX 模板的 MIStatIE 大佬，这份模板真的很漂亮；感谢我的导师**李斌斌**给予我充足的探索空间，提供给我优质的计算资源让我无限拓展能力的边界；感谢**吴大衍**老师、**秦绪功**老师科研上的指导和点拨；也要感谢我的爸爸妈妈，一直支持我自由地追寻热爱。💖

难得的周末，坐在五道口的 Page One 书店完成了整本笔记的最后部分。马上就要跨年了，隔着窗户望着外面熙熙攘攘的人群，回想当初整理这份笔记的缘由，也不知道是哪来的勇气！但转念又一想，还是给这一年留下了很多痕迹，意义非凡...🤓

新年快乐、后会有期、江湖再见！👋

杨桂淼

2025 年 12 月

北京 • 五道口

Part VII

附录

第 A 章

符号与记号（示例）

- $\mathbb{E}$：数学期望
- Var：方差算子

参考文献

[1] Vershynin, R. *High-Dimensional Probability*. CUP, 2018.