Department of Computer Science and Technology, Tsinghua University

计算语言学 HW5：歌词生成

刘泓尊　2022210866　计算机系

2022 年 11 月 27 日

1 Methods

利用课程提供的歌词语料，基于 GPT-2 架构，训练了 3 个歌词生成模型，训练环境为 1 张 NVIDIA 1080Ti (12GB)。模型细节如下：

1. **从头预训练 GPT-2** (GPT2)。

   参数配置：最大句子长度 =320、词表大小 =21128、向量维数 =768、12 层 Decoder、每层 12 个注意力头。backbone 参数量 124M.

   预训练配置 Epoch=20, Batch Size=32, 学习率 `1e-3`，Adam 优化器 ($\epsilon = 10^{-7}, \beta = (0.9, 0.98), weight\_decay = 10^{-3}$), 梯度裁剪到 1.0. warmup 6000 步。

   生成配置使用 Beam Search Multinomial Sampling. `num_beams=4, temprature=0.9, top_k=40, top_p=0.9, repetition_penalty=1.5`

2. **fine-tune** `uer/gpt2-chinese-cluecorpussmall` (GPT2$_{finetune}$)。

   该模型基于CLUECorpusSmall预训练。**预训练的参数配置**为: 最大句子长度 =1024、词表大小 =21128、向量维数 =768、12 层 Decoder、每层 12 个注意力头。backbone 参数量 124M.

   fine-tune 配置 Epoch=5, Batch Size=32, 学习率 `2e-5`，Adam 优化器 ($\epsilon = 10^{-7}, \beta = (0.9, 0.98), weight\_decay = 10^{-3}$), 梯度裁剪到 1.0. warmup 6000 步。

   生成配置使用 Beam Search Multinomial Sampling. `num_beams=4, temprature=0.9, top_k=40, top_p=0.9, repetition_penalty=1.5`

3. **fine-tune** `uer/gpt2-distil-chinese-cluecorpussmall` (GPT2-distil$_{finetune}$)。

   该模型基于CLUECorpusSmall预训练，并使用`uer/gpt2-chinese-cluecorpussmall`做知识蒸馏。**预训练的参数配置**为: 最大句子长度 =1024、词表大小 =21128、向量维数 =768、6 层 Decoder、每层 12 个注意力头。backbone 参数量 82M.

   fine-tune 配置 Epoch=5, Batch Size=32, 学习率 `2e-5`，Adam 优化器 ($\epsilon = 10^{-7}, \beta = (0.9, 0.98), weight\_decay = 10^{-3}$), 梯度裁剪到 1.0. warmup 6000 步。

生成配置使用 Beam Search Multinomial Sampling. `num_beams=4, temprature=0.9, top_k=40, top_p=0.9, repetition_penalty=1.5`

2 Experiment

使用 PPL 和 MAUVE 衡量测试集上的文本生成质量（在验证集 MAUVE 最大时进行测试）。计算 MAUVE 所用的特征提取网络为gpt2-medium，其他参数默认。模型自由生成与测试集实例相同数量的文本，计算 MAUVE 分数；PPL 为测试集上的 exp(*EnrtopyLoss*)。

2.1 Results on Different Models

表 1: Metrics on test set using different models using params descripted in Sec 1.

Models	PPL↓	MAUVE↑
GPT2	6.434	**0.475**
$\text{GPT2}_{finetune}$	**6.175**	0.459
$\text{GPT2-distil}_{finetune}$	6.456	0.466

三个模型的表现很接近。$\text{GPT2}_{finetune}$ 的 PPL 最佳，GPT2 的 MAUVE 最高。

2.2 Results on Different Params

针对 $\text{GPT2}_{finetune}$ 做了调参。
包括训练学习率、生成 `top_k, top_p, repetition_penlty`。其余参数同1小节。

2.2.1 Effects of Learning Rate

表 2: Metrics on test set using different learning rate.

Learning Rate	PPL↓	MAUVE↑
`1e-5`	6.183	**0.461**
`2e-5`	**6.175**	0.459
`5e-5`	6.198	0.453

对预训练模型的 Fine-tune 常常使用 `1e-5` 数量级的学习率，表2对比了不同学习率下的模型性能，可以看到 `1e-5` `2e-5` 都是比较适合的学习率，两者在 PPL 和 MAUVE 分数上各有优势。我们选择 `2e-5` 作为最终学习率。

2.2.2 Effects of `top_k`

Top-k 采样每次选择概率最高的 k 个样本，经过归一化之后再次采样。Top-k 结合 Beam Search 可以避免生成的句子陷入局部最优。表3对比了 $\text{GPT2}_{finetune}$ 模型不同 topk 下的测试集性能。可以看到 `top_k = 40` 时性能最佳，因为此时模型的生成有更多的选择。我们最终选择 `top_k = 40`。

表 3: Metrics on test set using different `top_k`.

`top_k`	PPL↓	MAUVE↑
2	6.176	0.018
10	6.176	0.459
40	**6.175**	**0.459**

2.2.3 Effects of `top_p`

表 4: Metrics on test set using different `top_p`.

`top_p`	PPL↓	MAUVE↑
0.1	6.176	**0.498**
0.5	6.176	0.452
0.9	**6.175**	0.459

Top-p(nucleus) 采样每次选择概率累加和恰好大于 p 的若干个 token，经过归一化之后再次采样。Top-p 结合 Beam Search 可以避免生成的句子陷入局部最优。表4对比了 $\text{GPT2}_{finetune}$ 模型不同 topp 下的测试集性能。可以看到 `top_p = 0.1,0.5,0.9` 时性能差距不大。我们最终选择 `top_p = 0.9`。

2.2.4 Effects of `repetition_penlty`

表 5: Metrics on test set using different `repetition_penalty`.

`repetition_penalty`	PPL↓	MAUVE↑
0.1	6.176	0.215
1.5	**6.175**	**0.459**
10.0	6.176	0.367

Repetition Penalty 可以被用来惩罚重复的词，定义为生成过程中下一个 token 在已生成序列中的重复次数，该损失的系数越大，模型生成结果重复性越低。表5对比了 $\text{GPT2}_{finetune}$ 模型不同 `repetition_penalty` 下的测试集性能。可以看到 `repetition_penalty = 1.5` 时性能最佳，过低的系数会使得重复 token 变多，过高的系数会使得生成时忽视模型的 language modeling 能力。我们最终选择 `repetition_penalty = 1.5`。

3 Examples

给定了 8 个经典歌词开头. 三个模型的生成结果见附录A，最大生成长度为 320。生成参数同 Sec 1.

3.1 Evaluation

可以看到每个模型生成时第一个字都有很大的重复性，对于 GPT2$_{finetune}$，8 句中有 6 句第一个 token 是“绻”；对于 GPT2-distil$_{finetune}$，8 句中有 3 句第一个 token 是“吸”；此外，我们还可以注意到生成的句子基本上主题和给的开头一致，比如开头“仁慈的父，我已坠入，看不见罪的国度。请原谅我，我的自负，”对应的生成结果中有“蜷缩在黑暗的角落里，没有任何人，能够将我救赎”(GPT2$_{finetune}$) 和“我们曾经拥有的一切，只是一片荒芜的土地，而我依然在这里，”(GPT2-distil$_{finetune}$)。这种现象可能是因为 GPT2 对于句子主题或风格的 modeling 能力比较强，但是却总是选择训练集中频率较高的 token 作为第一个 token 进行生成。

此外，三个模型对长句子的生成能力都不是很好。在长度超过 200 个词之后，三个模型的生成结果都会大量出现“爱情”、“生命”等实体，这可能是因为 pretrain 和 finetune 的数据集中都有大量关于“爱情”的段落。当生成句子过长时，GPT-2 也不能注意到所给开头的主题，这可能是因为 language modeling 的优化目标和自回归的生成方式导致的，序列的生成依然关注局部。

总的来说，三个模型在所给数据集上都有不错的生成结果，表现在：(1) 生成的子句长度都很短，一般为 4-8 个字符，很符合歌词的长度分布; (2) 续写歌词的主题和所给开头基本相同，尤其是关键词的情感倾向很接近；(3) 生成的句子存在押韵，符合歌词的特征。（比如 GPT2-distil$_{finetune}$ 的“吸着空气，想念着你，那么甜蜜，是否还记得，我们曾经说过的誓言，如今只剩下回忆，不知道何时才能忘记，也许明天会更美丽”）。

对比三个模型，GPT2-distil$_{finetune}$ 的生成结果更接近人类，押韵现象很多，生成的子句也具备多样性；而 GPT2$_{finetune}$ 生成的相似性更高，这可能是因为模型参数量过大，发生了过拟合；GPT2 由于没有进行预训练，其 diversity 更低，生成的句子押韵现象较少，续写文本和给定开头的关联性弱，甚至出现了“萨克斯特洛夫斯基隆冬子铭”这一不通顺的语句，说明没有经过大规模预训练的模型生成的文本更欠逻辑。

A Generations

Prefixes

1 仁慈的父，我已坠入，看不见罪的国度。请原谅我，我的自负，
2 我一路向北，离开有你的季节，你说你好累，已无法再爱上谁，
3 我给你的爱写在西元前，深埋在美索不达米亚平原。
4 对这个世界，如果你有太多的抱怨，跌倒了，就不敢继续往前走，
5 为你弹奏肖邦的夜曲，纪念我死去的爱情，跟夜风一样的声音，心碎的很好听，
6 塞纳河畔，左岸的咖啡，我手一杯，品尝你的美，留下唇印的嘴，
7 我送你离开，千里之外，你无声黑白。沉默年代，或许不该，太遥远的相爱。
8 天青色等烟雨，而我在等你。炊烟袅袅升起，隔江千万里。

Generation of **GPT2**

1 [仁慈的父，我已坠入，看不见罪的国度。请原谅我，我的自负，]你曾经说过，一定要坚持到最后，但是现在，还没有开始，就已结束，所有的错误，都被你狠地埋伏，而我只能，躲在黑暗里哭泣，等待着你的消息，为什么你从不怀疑，这样的爱情，让我无法再继续下去，那些伤痛的过程，使我变得如此冷静，当你转身离去的时候，眼中充满了泪水，你说你要走，因为我并不能够停止

2 [我一路向北，离开有你的季节，你说你好累，已无法再爱上谁，]滋润了心扉，没有人能够体会，只能默流下眼泪，也许是我太愚昧，才让你伤心欲绝，不知道该如何面对，这样的结果，让我如此的狼狈，曾经的承诺，都化作了灰飞烟灭，为你付出那么多，却换不来你给的一切，到底是谁犯了错，还是你根本就不在乎我，既然你已经远走，何必再去强求什么，当初的誓言，早已被现实所掩埋

3 [我给你的爱写在西元前，深埋在美索不达米亚平原。]萨克斯特洛夫斯基隆冬子铭，这是一个充满诗意的夏天，我们相约在巴黎铁塔下，听到了钟声响起的时候，那是多么美丽动人的夜晚，当我骑着单车穿过大街小巷，去寻找自己想要去的地方，没有人知道我的忧伤，只有那些被遗忘的角落里，藏着一些许多未完成的梦想，也许有一天我会离开这里回来，亲吻着石碑旁坐的木雕花窗

4 [对这个世界，如果你有太多的抱怨，跌倒了，就不敢继续往前走，]畏惧着什么，也许只是一种习惯，我想要的生活，总是那么的简单，却又不知该怎么办，害怕被冷落，所以才会更难过，没有人能告诉我，到底还要多久，才能找到出口，其实我并不懦弱，只希望自己能够，好的活下去，别再为失败而伤心，因为我们都在努力，用尽了全部力气，坚强的面对，未来的路途充满艰辛

5 [为你弹奏肖邦的夜曲，纪念我死去的爱情，跟夜风一样的声音，心碎的很好听，]你就这样离去，消失在我的世界里，我不知道该怎么做，才能让你回到我身边，给我最后一次机会，如果没有遇见你，请你不要再哭泣，告诉我你还爱着我，为什么你那么的冷漠，让我受尽了折磨，难道你真的忘了我吗，当初是你把我伤害了，留下我一个人生活，也许是我对你太过依赖，从此以后你是否会回来

6 [塞纳河畔，左岸的咖啡，我手一杯，品尝你的美，留下唇印的嘴，]滋润我的心扉，那是谁的眼泪，在爱情里徘徊，多少次梦中相会，醒来却不知疲惫，曾经的誓言，已化成了灰，难道这就是宿命的轮回，当初的离别，让我如何面对，为什么爱到最后只剩下伤痕累，看着你远去的背影，忍不住的泪水，思念的滋味，独自承受伤悲，想起你说的话语，深刺痛我心底，永远不后悔

7 [我送你离开，千里之外，你无声黑白。沉默年代，或许不该，太遥远的相爱。]萨克斯，夜晚的月光，像一个男人的眼睛，那样的皎洁，让我想起，曾经有过的温暖，如今却变得如此苍白，我的爱情是否还存在，也许你早已把我忘怀，只留下空荡的房间，弥漫着寂寞的味道，没有你的日子，真的好孤单，因为我知道你会回来，亲爱的姑娘，请你原谅我，从此不会再让你流泪，其实我并不是不懂你

8 [天青色等烟雨，而我在等你。炊烟袅袅升起，隔江千万里。]儿时的回忆，是否还惦记，年少的时光，我们曾一起努力，追逐过的梦想，如今已成为过往，岁月的痕迹，依然清晰，当年的烟火，燃烧成灰烬，那些曾经的美好，都随风远去，繁华落尽，却不见踪影，只剩下残缺的星，照亮了黑夜的心，多想再次看到你，微笑的样子，熟悉的面孔，陌生的表情，渐行渐远，模了眼神的脸颊

Generation of **GPT2**$-finetune$

1 [仁慈的父，我已坠入，看不见罪的国度。请原谅我，我的自负，]蜷缩在黑暗的角落里，没有任何人，能够将我救赎，只有我自己，才能体会到，生命的意义，这是我唯一的权利，因为我知道，世界上还有很多事情，需要我去改变，所以我决定离开这个国家，回到最初的地方，就让我重新来过，从此我不再孤独，虽然我曾经迷失了方向，但我相信，总有一天我会成为，最美丽的长发姑娘

2 [我一路向北，离开有你的季节，你说你好累，已无法再爱上谁，]绝的泪水，滑落在我的脸庞，如果时间可以倒退，我愿意为你而流浪，就算天涯海角，也要陪着你到地老天荒，不管未来多么遥远，只要能和你在一起，永远都不会分离，因为我爱你，所以我相信，这一切都是命中注定，没有什么能够阻挡，让我们一直走下去，直到世界末日来临，牵手走过每一个孤单的夜里，谢你陪我度过

3 [我给你的爱写在西元前，深埋在美索不达米亚平原。]绝的阳光，照耀着我们的家园，我为你披上白色的嫁衣，这是一个充满诗意的季节，你是我生命中最美丽的女人，让我忘记所有的忧伤，只想和你一起到永远，无论走过多少个春秋冬夏，也不能把你忘怀，因为我已深爱上了你，如果没有遇见你就不会有今生的缘分，那么我愿意用一生陪你到老，虽然我很小心翼地对待

4 [对这个世界，如果你有太多的抱怨，跌倒了，就不敢继续往前走，]绝的阳光，洒在我的脸上，那是我最美的梦想，让我们一起去飞翔，相信自己会变得更坚强，因为有你在身旁，所以我不再害怕受伤，感谢你给我力量，陪我走过每一个地方，也许有一天，你会发现，原来生命中，还有很多美好的愿望，虽然没有人能够告诉我，但我知道你一定会，一直都在这里，面对着未来，只有努力和拼搏

5 [为你弹奏肖邦的夜曲，纪念我死去的爱情，跟夜风一样的声音，心碎的很好听，]绝的旋律，仿佛在诉说着我们的过去，如果时间可以倒流，我愿意和你永远在一起，不管未来有多少风雨，只要有你在我身边，就不会孤单寂寞，因为有了你我什么都不怕，陪你度过每一个春夏秋冬，这是我最后一次想你，也是唯一的决定，希望你能够幸福快乐，让我抱着你一直到白头，记得你曾对我说过，那是一生的美丽

6 [塞纳河畔，左岸的咖啡，我手一杯，品尝你的美，留下唇印的嘴，]绝相随，爱情的味道，让人陶醉，这就是我想要的幸福，不管时间多么漫长，只要有你在身边，世界就会变得很美好，如果有一天我们老了，也许还能牵着手，感受彼此的温柔，那些曾经说过的承诺，现在都变成了泡沫，所有的回忆都被风吹走，最后剩下我一个人，独自面对寂寞，守着孤单的角落，每当夜幕降临之时

7 [我送你离开，千里之外，你无声黑白。沉默年代，或许不该，太遥远的相爱。]绫罗绸缎，还有一双眼睛，那是我最爱的女人，在我心里，永远都不会忘记，这个世界，没有什么能够代替，我们曾经拥有的美丽，如今只剩下了回忆，也许这就是爱情，所谓的永恒，不过是一场梦而已，可是谁又能说清楚，为何要这样伤害自己，从此以后再也不能见到你，明知道你不爱我，偏还要把我捧在手心

8 [天青色等烟雨，而我在等你。炊烟袅升起，隔江千万里。]绝的月光，照亮了我的心，我知道你已经离去，只留下一片空虚，那是我最后一次想你，也是唯一的回忆，如果时间可以倒流，让我们重新开始，就不会再有这样的结局，为什么我总是不能忘记，你的美丽，因为我曾经爱过你，所以我才会选择放弃，把你放在我心里，从此我不再与你相依，当我回头看着你消逝

Generation of **GPT2-distil**−*finetune*

1 [仁慈的父，我已坠入，看不见罪的国度。请原谅我，我的自负，]我们曾经拥有的一切，只是一片荒芜的土地，而我依然在这里，守护着你，直到永远，无论我走到哪里都不会忘记，你的名字，那是我心中最美的回忆，如果有一天我真的离去，就让我再一次为你唱一首歌给你听，当我睁开双眼，才发现世界已变得如此的美丽，从此以后我将不再哭泣，因为我知道

2 [我一路向北，离开有你的季节，你说你好累，已无法再爱上谁，]吸着空气，想念着你，那么甜蜜，是否还记得，我们曾经说过的誓言，如今只剩下回忆，不知道何时才能忘记，也许明天会更美丽，就让我陪在你身边，给你我所有的温柔，为你唱一首歌，让你听见我的心跳，这样的夜晚我真的好孤单，没有人能够代替，因为我相信爱情可以永远，和你一起看日落，漫天的星，它是最美的梦

3 [我给你的爱写在西元前，深埋在美索不达米亚平原。]这是一个古老的传说，它象征着生命的奇迹，那是因为有了你，我们才能够相遇，从此以后，世界变得更加美丽，如果你愿意，就请你告诉我，你是否还会记得，曾经拥有过的一切，而如今却只剩下回忆，多么希望时间可以停止，让我再次回到你身边，也许你已经忘记了，最初的誓言，直到现在我才明白，有些故事早已不再有意思

4 [对这个世界，如果你有太多的抱怨，跌倒了，就不敢继续往前走，]吸着空气，感受着自己，生活的压力，我们都一样，没有什么能够阻挡，因为有你在身旁，每一天都是新的开始，所有的烦恼全部抛在脑后，让我来陪你唱首歌，告诉你我的梦想，虽然很遥远，但我会一直坚持到底，给你最好的答案，只要你愿意，永远都不放弃，爱情的路上，充满荆棘和坎坷，请你相信我，希望你也曾经失意过

5 [为你弹奏肖邦的夜曲，纪念我死去的爱情，跟夜风一样的声音，心碎的很好听，]也许这是最后的结局，让我不知所措，没有你的日子，我该怎么过，只想和你在一起，如果你愿意，就请你告诉我，你会爱我到底，还是我对你太过依赖，忘记了那些甜蜜的回忆，曾经说过的誓言，现在都已经随风而去，无法再继续，因为我们之间的距离，越来越远，留下我一个人独自伤心，为何你要离开

6 [塞纳河畔，左岸的咖啡，我手一杯，品尝你的美，留下唇印的嘴，]吸着爱的味道，看着你微笑的脸，幸福就在眼前，不知不觉，已经走到了终点，是否还记得，那个夏天，你牵着我的手，许下的诺言，从此不会改变，直到现在，只想和你永远，在一起，每一个夜晚，都有你在身边，让我为你写一首情歌，唱给你听，这一切，真的好甜蜜，没有人能够代替，如果你愿意，用心去爱我，请把它留在你背景下

7 [我送你离开，千里之外，你无声黑白。沉默年代，或许不该，太遥远的相爱。]希望你能明白，我的心在等待，期待你的出现，如果可以，让我为你唱一首歌，请你不要再犹豫，因为我知道，这一切都是命运的安排，只要你愿意，就会和我一起，永远不分离，所有的故事，都已经过去，每个人都有自己的主题，那些美好的回忆，全部都藏在心底，从此不再孤寂，直到有一天，世界都变成了海洋

8 [天青色等烟雨，而我在等你。炊烟袅升起，隔江千万里。]那是我的家乡，山青水秀的地方，一望无际的大草原，春风吹拂着她的脸庞，让我想起了远方的姑娘，如果有一天，我们不再相见，请把我带到你的身旁，啊，亲爱的朋友，你可知道我心中的忧伤，为什么这样的时候，总是难以忘怀，虽然我已经长大，却还是深爱着你，永远不会忘记，曾经的誓言，所有的往事都变得遥远