# 微调GPT-2的儿童医疗问答系统

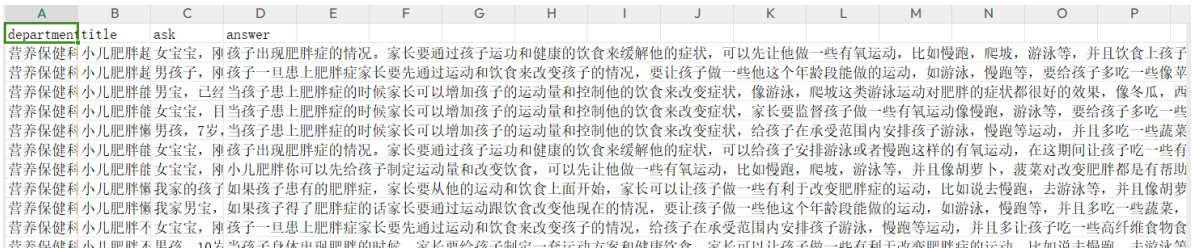


## 引言

**GPT-2（由OpenAI开发）是一个预训练的语言模型，设计用于处理各种文本生成任务。由于其强大的生成能力和广泛的训练数据，GPT-2 可以用于多种用途，从聊天机器人到内容生成，再到更为专业的领域，如儿童医疗问答。**

## 1.数据收集和准备

收集医药领域的问答数据集，本文选用的数据集来自于github的中文医学问答数据集项目，包含六个科室的医学问答数据，考虑到数据规模庞大，我们只选用了儿科的数据：



| department | title | ask | answer |
|---|---|---|---|
| 营养保健科 | 小儿肥胖超 | 女宝宝，刚 | 孩子出现肥胖症的情况。家长要通过孩子运动和健康的饮食来缓解他的症状，可以先让他做一些有氧运动，比如慢跑，爬坡，游泳等，并且饮食上孩子 |
| 营养保健科 | 小儿肥胖超 | 男孩子，刚 | 孩子一旦患上肥胖症家长要先通过运动和饮食来改变孩子的情况，要让孩子做一些他这个年龄段能做的运动，如游泳，慢跑等，要给孩子多吃一些像苹 |
| 营养保健科 | 小儿肥胖能 | 男宝，已经 | 当孩子患上肥胖症的时候家长可以增加孩子的运动量和控制他的饮食来改变症状，像游泳，爬坡这类游泳运动对肥胖的症状都很好的效果，像冬瓜，西 |
| 营养保健科 | 小儿肥胖能 | 女宝宝，目 | 当孩子患上肥胖症的时候家长可以增加孩子的运动量和控制他的饮食来改变症状，家长要监督孩子做一些有氧运动像慢跑，游泳等，要给孩子多吃一些 |
| 营养保健科 | 小儿肥胖懒 | 男孩，7岁， | 当孩子患上肥胖症的时候家长可以增加孩子的运动量和控制他的饮食来改变症状，给孩子在承受范围内安排孩子游泳，慢跑等运动，并且多吃一些蔬菜 |
| 营养保健科 | 小儿肥胖能 | 女宝宝，刚 | 孩子出现肥胖症的情况。家长要通过孩子运动和健康的饮食来缓解他的症状，可以给孩子安排游泳或者慢跑这样的有氧运动，在这期间让孩子吃一些有 |
| 营养保健科 | 小儿肥胖能 | 女宝宝，刚 | 小儿肥胖你可以先给孩子制定运动量和改变饮食，可以先让他做一些有氧运动，比如慢跑，爬坡，游泳等，并且像胡萝卜，菠菜对改变肥胖都是有帮助 |
| 营养保健科 | 小儿肥胖懒 | 我家的孩子 | 如果孩子患有的肥胖症，家长要从他的运动和饮食上面开始，家长可以让孩子做一些有利于改变肥胖症的运动，比如说去慢跑，去游泳等，并且像胡萝 |
| 营养保健科 | 小儿肥胖懒 | 我家男宝， | 如果孩子得了肥胖症的话家长要通过运动跟饮食改变他现在的情况，要让孩子做一些他这个年龄段能做的运动，如游泳，慢跑等，并且多吃一些蔬菜， |
| 营养保健科 | 小儿肥胖不 | 女宝宝，刚 | 孩子一旦患上肥胖症家长要先通过运动和饮食来改变孩子的情况，给孩子在承受范围内安排孩子游泳，慢跑等运动，并且多让孩子吃一些高纤维食物食 |
| 营养保健科 | 小儿肥胖不 | 男孩，10岁 | 当孩子身体出现肥胖的时候，家长要给孩子制定一套运动方案和健康饮食，家长可以让孩子做一些有利于改变肥胖症的运动，比如说去慢跑，去游泳等 |

**文件格式如上所示，一共有101603条数据。**

## 2.数据预处理

我们对数据进行处理，删除空数据，并根据gpt-2数据格式要求对“department”一特征进行删除，title与ask进行合并，得到最终符合gpt-2要求的数据格式。

```
#空数据删除

import pandas as pd

# 读取CSV文件，指定编码格式为gb18030
data = pd.read_csv('data/儿科5-14000.csv', encoding='gb18030')

# 获取原始数据的行数
original_rows = data.shape[0]

# 删除包含空值的行
```

```
data.dropna(inplace=True)

# 获取处理后数据的行数
cleaned_rows = data.shape[0]

# 保存修改后的数据到新的CSV文件，指定编码格式为gb18030
data.to_csv('cleaned_file.csv', index=False, encoding='gb18030')
```

```
# 数据转换

import json

def prepare_data(json_file, output_file):
    with open(json_file, 'r', encoding='utf-8') as file:
        data = json.load(file)

    prepared_data = []
    for item in data:
        question = item['question']
        answer = item['answer']
        prepared_data.append(f"<question>{question}<answer>{answer}")

    with open(output_file, 'w', encoding='utf-8') as file:
        file.write('\n'.join(prepared_data))

# 指定你的JSON文件路径和输出文件路径
json_file = 'output_modified.json'
output_file = 'prepared_data.txt'

# 转换并保存数据
prepare_data(json_file, output_file)
```

最终数据格式：

```
<question>宝宝长牙发烧呕吐拉稀需要吃哪些药？个多月病情描述:我小孩长牙齿3天了 发烧 呕吐 拉稀是自然症状吗曾经治疗情况和效果:不好反反复复想得到怎样的帮助:该吃什么药<answer>在牙齿努力钻出牙龈的时候,难免会出现类似"伤口"的地方,产生不适。如果此时口腔清洁度不够,牙龈发炎、发烧的可能性很大。因此,家长要多帮宝宝清洁口腔,平时多喝水,饭后注意漱口。如果体温超过38.5℃,应及时到医院就诊。长牙的同时,宝宝的唾液腺发育渐趋成熟。因此,多数宝宝会不停地流口水。家长要及时为其擦干口水,避免损伤局部皮肤。宝宝的上衣、枕头、被褥等容易被口水污染,要勤洗勤晒,以免滋生细菌。
<question>婴儿全身出现小红点并伴随发热症状的疑问？前天晚上宝宝再次出现发烫的症状，然后第三天早上去医院发觉喉咙再次出现了红肿并且伴随着发烫的症状，于是医生给宝宝挂了吊瓶和许多药，后来热就已经开始退了，但没多久又已经开始发烫了，今天发觉宝宝还再次出现了小红点。 在乎怎样的帮助，您好为什么婴儿会再次出现全身小红点并且有发烫的症状？<answer>您好：根据你的叙述，小儿的全身红点不退可能会和小儿发烧引来的疹子有一定的关系，建议你还是留意正确的护理小儿的方法。一般小儿的发烫可以是当室温过高、衣服穿的过多、炎症性疾病等情况可再次出现发烫，应及时就诊，查明原因。 以上是对"婴儿全身再次出现小红点并还伴发烫症状的疑虑"这个问题的建议，期望对您有帮助，祝您健康!
<question>4个半月打喷嚏流鼻子，偶尔咳嗽怎么办？好几天了，流鼻涕，打喷嚏，偶尔咳嗽，吃了两天护彤小儿黄娜敏颗粒也不管用该怎么办<answer>你这种情况可能是感冒引起的，需要口服感冒药、抗生素、抗过敏、抗病毒等药物治疗。平时需要加强运动、增强体质。禁吃辛辣酒、生冷等刺激性食物。，孩子被新生儿疾病所困扰，因此，家长一定要带孩子立即就医治疗，避免出现种种误区，要及时对症治疗，对于家长来说要给予孩子足够的关心和照顾，要尽可能的帮助孩子及时治疗疾病。
<question>孩子中耳炎耳朵痒该怎么样治较好？女宝宝，目前3岁，这几天，孩子的耳朵有点疼，另外，有黄色的耳屎流出，另外，好像没什么食欲也很乏力，请问：孩子中耳炎耳朵痒该怎么样治较好。<answer>可适当的使用一些抗生素或是消炎类药物，也可局部用药，比如给宝宝采取使用消炎类型的滴耳剂，孩子如果耳朵比较疼的话也适量情给一些镇痛的药物，另外如果伴有发烧的情况的话，那么也可服用一些退烧药，高烧的话还是建议要尽早就医的，以上都是比较常用的治疗方法，但是如果孩子出现了耳膜穿孔的症状，需要及时的去医院进行手术治疗，治疗的同时也要注意给孩子安排清淡的饮食，加强营养的补充，帮助尽早康复。
```

## 3.GPT-2模型的微调

### 数据准备

我们使用两个文本文件进行训练和测试（训练集比测试集为8:2）：

- `train.txt`：用于训练的数据文件。
- `test.txt`：用于评估模型性能的数据文件。

这些文件包含儿童医疗相关的问题和答案。

**模型和分词器的初始化**

首先，我们使用 `GPT2Tokenizer` 和 `GPT2LMHeadModel` 从预训练的 "gpt2" 模型初始化分词器和模型。

```
model_name = "gpt2"
tokenizer = GPT2Tokenizer.from_pretrained(model_name)
model = GPT2LMHeadModel.from_pretrained(model_name)
```

**数据集准备**

使用 `TextDataset` 类，我们根据上述文本文件和分词器创建了训练和测试数据集。

```
python
train_dataset = TextDataset(tokenizer=tokenizer, file_path=train_path,
block_size=512)
test_dataset = TextDataset(tokenizer=tokenizer, file_path=test_path,
block_size=512)
```

为了在训练时使用这些数据集，我们还创建了一个 `DataCollatorForLanguageModeling` 实例。

**评估指标**

我们定义了一个评估函数 `compute_metrics` 来计算模型的困惑度。困惑度是评估语言模型好坏的常用指标。

```
def compute_metrics(p: EvalPrediction):
    # 困惑度 (perplexity) 计算
    perplexity = torch.exp(torch.tensor(p.predictions).mean()).item()
    return {"perplexity": perplexity}
```

**训练参数**

为了微调模型，我们设置了以下训练参数：

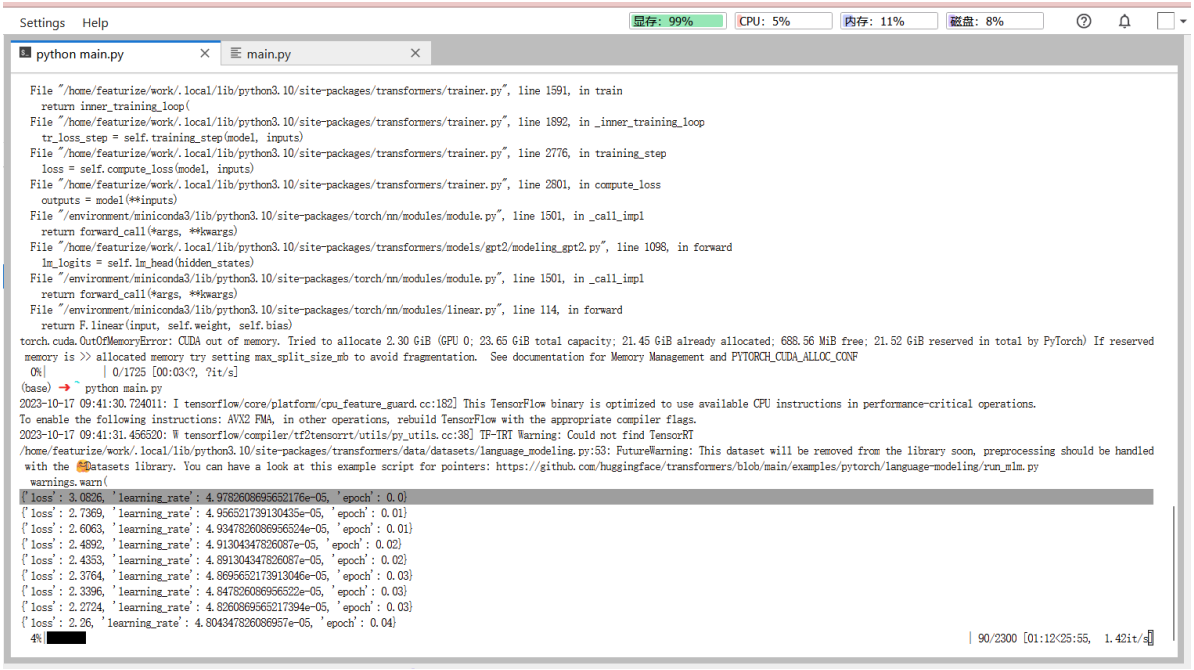
- Batch size
- 梯度累积步数
- 训练轮数
- 日志和保存的步数
- 评估策略

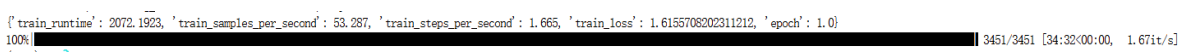
```
training_args = TrainingArguments(
    per_device_train_batch_size=24,
    per_device_eval_batch_size=24,
    gradient_accumulation_steps=2,  # 使用梯度累积
    num_train_epochs=1,
    logging_dir='./logs',
    logging_steps=10,
    save_steps=10,
    eval_steps=10,
    evaluation_strategy="epoch",
    save_total_limit=2,  # 最多保存的模型数量
    output_dir="./gpt2_finetuned_model"
)
```

## 4. 训练和评估

使用 `Trainer` 类，我们传入上述所有的组件（模型、数据集、训练参数等）并开始训练。完成训练后，我们进行了模型评估。

训练配置：


**loss值总体来看在不断变小，学习率大概在 10-6与10-5之间**

微调后的模型与分类器保存在：gpt2_finetuned_model


## 5. 结果

训练后，模型应在测试数据集上显示其困惑度。这个指标可以帮助我们了解模型在儿童医疗问答任务上的性能。

最后：

**perplexity： 18**

说明我们的模型效果不是很好，可能需要一些改进，我们综合考虑，得到以下几种改进思路:

1.增加训练数据量：更大的训练数据集通常可以提高模型的泛化能力，减少过拟合的可能性。我们的训练数据为8w条左右，对于gpt-2这样的模型来说可能不够多；

2.数据预处理和输入表示：确保进行适当的数据预处理，例如标记化、去除噪声、规范化等。此外，考虑使用更好的输入表示方法，如词嵌入技术，以提高模型的表达能力。

## 6.实例演示

```
35 # 3. 使用函数
36 question = "小儿发烧时高时低怎么办？"
37 print(ask_question(question))
```

```
(base) ➜ python run.py
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
小儿童发烧，家长先可酒精擦浴退热，然后医院查有无炎症，先不急着用药。
```

```
# 3. 使用函数
question = "打完乙肝疫苗能吃增强免疫力的药吗？"
print(ask_question(question))
```

```
(base) ➜ python run.py
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
小儿肝病病因多样，一旦确诊，需要及时治疗，家长要配合医生。要注意肠胃减压和良好的卫生习惯。孩子饮食要合理，避免寒冷刺激的食物
```

总体来说，模型能够回答一定的问题，但是有时候语句不是很流畅，而且有时候会误解问题重点。

## 7. 结论

微调 GPT-2 用于儿童医疗问答是一个有前景的方向。虽然原始的 GPT-2 已经在多种任务上表现出色，但通过专门的微调，我们可以使其更适应特定的应用，如儿童医疗问答。我们期望这种模型能为医疗专家提供支持，帮助他们更有效地回答关于儿童健康的问题。