



第四次实践课作业-Transformer中英翻译

罗浩铭 PB21030838

题目

修改课程中的代码，使其完成中英翻译（原代码是英中翻译）同时，将encoder layer的个数设为4，将decoder layer的个数设为5 展示训练后的中英翻译结果

实验内容

实现中英翻译

修改encoder与decoder的个数

我们维持API规范来进行修改

先修改配置信息：

```
N_ENCODER = 4
N_DECODER = 5
```

同时修改 `make_model` 函数：

先修改函数的参数 `N` 为 `n_encoder, n_decoder`：

```
def make_model(src_vocab, tgt_vocab, n_encoder = N_ENCODER,
               n_decoder = N_DECODER, d_model=512, d_ff=2048, h=8,
               dropout=0.1):
```

再将上述参数用于 `Encoder` 与 `Decoder` 的调用：

```

Encoder(EncoderLayer(d_model, c(attn), c(ff),
                        dropout).to(DEVICE),
n_encoder).to(DEVICE)
Decoder(DecoderLayer(d_model, c(attn), c(attn),
                        c(ff), dropout).to(DEVICE),
n_decoder).to(DEVICE)

```

再按上述API修改 `make_model` 的调用：

```

# Step 2: Init model
model = make_model(
    src_vocab,
    tgt_vocab,
    N_ENCODER,
    N_DECODER,
    D_MODEL,
    D_FF,
    H_NUM,
    DROPOUT
)

```

由此完成修改。

debug

DEVICE类型判断条件错误

源代码中存在如下的判断语句：

```

if DEVICE == "cuda":

```

但由于 `DEVICE = torch.device("cuda")`，因此 `DEVICE` 的类型是 `torch.device`，而不是 `str`，因此上述判断语句永远为 `False`，与预期值 `True` 不符。

因此，需要将上述判断语句修改为：

```
if DEVICE.__str__() == "cuda":
```

实验结果

实验组别	best-loss
baseline	1.788
pretrained	0.834

附：练习

0. 看看原始数据长什么样？

略

1. 把代码跑通（所有的图都是可以展示的，对于理解Transformer非常有用）（含Debug和非Debug）

略

1.5 试试预训练模型 save/models/large_model.pt，与你自己训的对比一下

Baseline翻译测试结果如下：

```

BOS i 've changed my website 's layout . EOS
BOS 我改了一下我的网站的布局。EOS
translation: 我用了我的火車站要花了。

BOS he brought back several UNK . EOS
BOS 他帶回了一些紀念品。EOS
translation: 他按下了一些大的多少。

BOS he is very angry . EOS
BOS 他非常生气。EOS
translation: 他非常生氣。

BOS how 's the weather in new york ? EOS
BOS 紐約的天氣如何？EOS
translation: 紐約的火車怎麼樣？

BOS let 's go to a movie . EOS
BOS 讓我們去看電影。EOS
translation: 讓我們一個電影。

BOS come on ! give me a chance . EOS
BOS 來嘛！給我個機會。EOS
translation: 你在一點上了我一個好的時間。

BOS she is a good UNK . EOS
BOS 她是一個很好的舞者。EOS
translation: 她是一個好的。

BOS some of them seem to be too difficult . EOS
BOS 其中一些似乎太难了。EOS
translation: 他們看起來很好。

BOS my parents usually speak to each other in french , even though my mother is a native english speaker . EOS
BOS 我父母通常用法語對話，即使我母親的母語是英語。EOS
translation: 我父親法語言，但我父母語言成為我父母語言老師。

BOS hurry up , or you will miss the train . EOS
BOS 快点，不然你就要错过火车了。EOS
translation: 快点你会错过火车。
<<<<<< finished evaluate, cost 7.1324 seconds

```

预训练模型翻译测试结果如下：

```

BOS i 've changed my website 's layout . EOS
BOS 我改了一下我的网站的布局。 EOS
translation: 我改了我的收音机器。

BOS he brought back several UNK . EOS
BOS 他带回了一些纪念品。 EOS
translation: 他到了按钮。

BOS he is very angry . EOS
BOS 他非常生气。 EOS
translation: 他非常生气。

BOS how 's the weather in new york ? EOS
BOS 纽约的天气如何? EOS
translation: 纽约天气怎么长椅子?

BOS let 's go to a movie . EOS
BOS 让我们去看电影。 EOS
translation: 我们去看电影。

BOS come on ! give me a chance . EOS
BOS 来嘛! 给我个机会。 EOS
translation: 来给我一点了机会。

BOS she is a good UNK . EOS
BOS 她是一个很好的舞者。 EOS
translation: 她是好的。

BOS some of them seem to be too difficult . EOS
BOS 其中一些似乎太难了。 EOS
translation: 他们似乎对它来说很难。

BOS my parents usually speak to each other in french , even though my mother is a native english speaker . EOS
BOS 我父母通常用法语对话, 即使我母亲的母语是英语。 EOS
translation: 我妈通常和我妈法语是一名的母语说英语。

BOS hurry up , or you will miss the train . EOS
BOS 快点, 不然你就要错过火车了。 EOS
translation: 快点, 你会错过一下火车。
<<<<<< finished evaluate, cost 8.9434 seconds

```

二者的翻译效果不相上下。

不过, Baseline的best-loss是1.788, 预训练模型的best-loss是0.834, 预训练模型的loss更低。

2. 看一看建立的英文词典和中文词典长什么样, 思考一下为什么要有 "UNK" "BOS" "EOS"

建立的英文词典和中文词典如下:

5277, 'bowl': 5278, 'murmur': 5279, 'brook': 5280, 'lulls': 5281, 'tragic': 5282, 'overnight': 5283, 'necessity': 5284, 'religious': 5285, 'abate': 5286, 'senator': 5287, 'censured': 5288, 'congressional': 5289, 'ethics': 5290, 'exists': 5291, 'pochi': 5292, 'span': 5293, 'butterfly': 5294, 'candlelight': 5295, 'mastering': 5296, 'influenced': 5297, 'delays': 5298, 'commanded': 5299, 'fork': 5300, 'umpire': 5301, 'dollar': 5302, 'homes': 5303, 'cardboard': 5304, 'pollen': 5305, 'birmingham': 5306, 'platform': 5307, 'justice': 5308, 'prevail': 5309, 'p.m.': 5310, 'sadiy': 5311, 'appeal': 5312, 'madrid': 5313, 'household': 5314, 'patriotism': 5315, 'remote': 5316, 'left-handed': 5317, 'physician': 5318, 'worm': 5319, 'shirtless': 5320, 'awfully': 5321, 'shed': 5322, 'cap': 5323, 'astronaut': 5324, 'toyota': 5325, 'ford': 5326, 'easy-going': 5327, 'fate': 5328, 'imposing': 5329, 'beliefs': 5330, 'cramped': 5331, 'weigh': 5332, 'female': 5333, 'capitals': 5334, 'publication': 5335, 'timed': 5336, 'coincide': 5337, 'seventy': 5338, 'misses': 5339, 'ambiguities': 5340, 'senses': 5341, 'slamming': 5342, 'thoughtful': 5343, 'microwave': 5344, 'comics': 5345, 'toothpaste': 5346, 'engagement': 5347, 'madman': 5348, 'frequent': 5349, 'progressing': 5350, 'abstained': 5351, 'invite': 5352, 'shoebox': 5353, 'capable': 5354, '40': 5355, 'scarf': 5356, 'interrupted': 5357, 'jet': 5358, 'landed': 5359, 'fourth': 5360, 'rats': 5361, 'simply': 5362, 'divine': 5363, 'spilt': 5364, 'wasting': 5365, 'underwater': 5366, 'survivors': 5367, 'topic': 5368, 'poker': 5369, 'drum': 5370, 'self-service': 5371, 'in-laws': 5372, 'bridges': 5373, 'ten-thirty': 5374, 'narrowly': 5375, 'handing': 5376, 'blocked': 5377, 'landslide': 5378, 'fireplace': 5379, 'developed': 5380, 'tool': 5381, 'drugstore': 5382, 'brazen': 5383, 'defiance': 5384, 'perform': 5385, 'educated': 5386, 'stroganoff': 5387, 'containers': 5388, 'airtight': 5389, 'universities': 5390, 'fast-moving': 5391, 'policemen': 5392, 'intruded': 5393, '1,600': 5394, 'retiring': 5395, '19th': 5396, 'loaves': 5397, 'alice': 5398, 'bible': 5399, 'priest': 5400, 'timetable': 5401, 'unlock': 5402, 'bribes': 5403, 'highway': 5404, '58': 5405, 'admired': 5406, 'non-smoking': 5407, 'violated': 5408, 'sour': 5409, 'stocks': 5410, 'germanic': 5411, 'lad': 5412, 'stack': 5413, 'political': 5414, 'career': 5415, 'capture': 5416, 'over-sleeping': 5417, 'spared': 5418, 'pains': 5419, 'annoyed': 5420, 'noises': 5421, 'rugby': 5422, 'differs': 5423, 'washington': 5424, 'drowsy': 5425, 'awarded': 5426, 'online': 5427, 'utterly': 5428, 'joints': 5429, 'extremes': 5430, 'satisfactory': 5431, 'accomplished': 5432, 'deposit': 5433, 'extreme': 5434, 'remind': 5435, 'nagasaki': 5436, 'youngsters': 5437, 'thirty-one': 5438, 'liking': 5439, 'strict': 5440, 'catching': 5441, 'cycling': 5442, 'short-sleeved': 5443, 'shirts': 5444, 'sailing': 5445, 'weekends': 5446, 'confidential': 5447, 'conceal': 5448, 'disappointment': 5449, 'proceed': 5450, 'agenda': 5451, '56': 5452, 'russell': 5453, 'welfare': 5454, 'sewage': 5455, 'pollutes': 5456, 'loser': 5457, 'statues': 5458, 'stopover': 5459, 'warrior': 5460, 'weakness': 5461, 'stage': 5462, 'roller': 5463, 'skates': 5464, 'occurred': 5465, 'deciding': 5466, 'glaive': 5467, 'accomplish': 5468, 'spilling': 5469, 'phrase': 5470, 'wealthy': 5471, 'modest': 5472, 'hawk': 5473, 'mouse': 5474, 'ceases': 5475, 'parrot': 5476, 'pat': 5477, 'cupcakes': 5478, 'ties': 5479, 'boarding': 5480, 'execution': 5481, 'eleventh': 5482, 'artist': 5483, 'calculated': 5484, 'opposition': 5485, 'springs': 5486, 'schoolroom': 5487, 'absence': 5488, 'fonder': 5489, 'field': 5490, 'educational': 5491, 'foster': 5492, 'UNK': 0, 'PAD': 1
['BOS': 2, 'EOS': 3, '': 4, '我': 5, '的': 6, '了': 7, '你': 8, '他': 9, '是': 10, '一': 11, '在': 12, '不': 13, '？': 14, '有': 15, '她': 16, '这': 17, '很': 18, '妈': 19, '，': 20, '要': 21, '人': 22, '天': 23, '们': 24, '个': 25, '去': 26, '到': 27, '这': 28, '场': 29, '上': 30, '们': 31, '个': 32, '什': 33, '好': 34, '想': 35, '那': 36, '吗': 37, '子': 38, '能': 39, '么': 40, '看': 41, '得': 42, '做': 43, '曾': 44, '多': 45, '下': 46, '生': 47, '吗': 48, '以': 49, '裡': 50, '大': 51, '来': 52, '可': 53, '度': 54, '会': 55, '道': 56, '都': 57, '来': 58, '没': 59, '喜': 60, '把': 61, '家': 62, '知': 63, '？': 64, '事': 65, '就': 66, '起': 67, '说': 68, '本': 69, '作': 70, '没': 71, '时': 72, '为': 73, '常': 74, '里': 75, '車': 76, '为': 77, '意': 78, '打': 79, '年': 80, '和': 81, '時': 82, '地': 83, '自': 84, '說': 85, '吃': 86, '用': 87, '成': 88, '出': 89, '工': 90, '日': 91, '些': 92, '法': 93, '孩': 94, '給': 95, '學': 96, '明': 97, '晚': 98, '點': 99, '前': 100, '今': 101, '最': 102, '英': 103, '開': 104, '小': 105, '中': 106, '信': 107, '請': 108, '学': 109, '回': 110, '哪': 111, '候': 112, '行': 113, '真': 114, '所': 115, '歡': 116, '房': 117, '對': 118, '告': 119, '它': 120, '跟': 121, '給': 122, '過': 123, '点': 124, '被': 125, '比': 126, '率': 127, '每': 128, '語': 129, '正': 130, '昨': 131, '早': 132, '何': 133, '欢': 134, '書': 135, '相': 136, '果': 137, '过': 138, '父': 139, '高': 140, '必': 141, '太': 142, '着': 143, '快': 144, '对': 145, '儿': 146, '心': 147, '公': 148, '您': 149, '加': 150, '再': 151, '手': 152, '开': 153, '已': 154, '住': 155, '该': 156, '非': 157, '美': 158, '湯': 159, '球': 160, '著': 161, '定': 162, '丽': 163, '感': 164, '談': 165, '0': 166, '花': 167, '十': 168, '走': 169, '間': 170, 'o': 171, '怎': 172, '名': 173, '西': 174, '玛': 175, '女': 176, '让': 177, '望': 178, '讓': 179, '經': 180, '友': 181, '老': 182, '面': 183, '间': 184, '语': 185, '話': 186, '发': 187, '需': 188, '完': 189, '见': 190, '著': 191, '间': 192, '话': 193, '火': 194, '次': 195, '應': 196, '親': 197, '只': 198, '男': 199, '母': 200, '力': 201, '书': 202, '見': 203, '己': 204, '也': 205, '現': 206, '明': 207, '電': 208, '問': 209, '物': 210, '认': 211, '朋': 212, '加': 213, 'T': 214, '还': 215, '士': 216, '從': 217, '三': 218, '之': 219, '同': 220, '！': 221, '照': 222, '現': 223, '音': 224, 'm': 225, '样': 226, '受': 227, '找': 228, '新': 229, '外': 230, '方': 231, '国': 232, '理': 233, '分': 234, '潜': 235, '少': 236, '但': 237, '等': 238, '水': 239, '雨': 240, '放': 241, '从': 242, '發': 243, '始': 244, '字': 245, '更': 246, '解': 247, '校': 248, '國': 249, '潜': 250, '東': 251, '情': 252, '希': 253, '星': 254, '長': 255, '吧': 256, '全': 257, '像': 258, '复': 259, '觉': 260, '經': 261, '認': 262, '应': 263, '聽': 264, '直': 265, '後': 266, '狗': 267, '待': 268, '电': 269, '重': 270, '於': 271, '無': 272, '別': 273, '餐': 274, '睡': 275, '喝': 276, '请': 277, '午': 278, '文': 279, 'I': 280, '服': 281, '玩': 282, '钱': 283, '月': 284,

其为python的 dict 对象，key为词，value为词的编号。其中，UNK 表示未知词，编号为0；PAD 表示填充词，编号为1；其余词为数据集分词后出现过的词，按出现频率从高到低编号为2~N。两个词典中，BOS 表示句子的开始，编号均为2；EOS 表示句子的结束，编号均为3。

UNK、PAD、BOS、EOS 的作用如下：

- UNK：未知词，当输入不在训练集里时，模型会遇到未知词时，这时需要将其替换为 UNK，以使得模型可以正常运行。
- PAD：填充词，句子长度小于模型接收的最大token长度时，用于填充剩余部分。
- BOS：句子开始，BOS 用于提示模型开始翻译一个句子。
- EOS：句子结束，EOS 用于提示模型结束翻译一个句子。

3. word embedding后的词长什么样？

word embedding由 Embeddings 类实现，代码如下：

```

class Embeddings(nn.Module):
    def __init__(self, d_model, vocab):
        super(Embeddings, self).__init__()
        self.lut = nn.Embedding(vocab, d_model)
        self.d_model = d_model

    def forward(self, x):
        # return x's embedding vector (times math.sqrt(d_model))
        return self.lut(x) * math.sqrt(self.d_model)

```

它使用了一个 `nn.Embedding` 对象，类似于查找表，输入每个词的编号即可得到每个该词的embedding，其为一个 `d_model = D_MODEL = 256` 维的向量。

因此，word embedding后每个词都是一个 `d_model = D_MODEL = 256` 维的向量。

4. 思考一下，为什么要加positional_embedding

这是为了让模型能够获得词的位置信息。因为Transformer架构中输入的各个位置是同时进行处理，在微结构上都是对等的，互相之间没有顺序关系，所以需要额外的方式来让模型学习到词的位置信息。