Homework3实验报告

姓名: 梁辉

学号: 10185501411

一、实验内容

实验三: 机器翻译

一、数据集

- 地址: http://www.statmt.org/wmt16/translation-task.html (WMT16)
- 选择新闻领域数据集:

Shared Task: Machine Translation of News

[HOME] [SCHEDULE] [PAPERS] [AUTHORS] [RESULTS]

TRANSLATION TASKS: [NEWS] [IT-DOMAIN] [BIOMEDICAL] [MULTIMODAL] [PRONOUN]

EVALUATION TASKS: [METRICS] [QUALITY ESTIMATION] [TUNING]

OTHER TASKS: [AUTOMATIC POST-EDITING] [BILINGUAL DOCUMENT ALIGNMENT]

• 选择英语->德语的翻译任务:

 $The \ recurring \ translation \ task \ of \ the \ \underline{WMT \ workshops} \ focuses \ on \ news \ text \ and \ European \ language \ pairs. For \ 2016 \ the \ language \ pairs \ are:$

实现英文到德语的翻译任务,并在测试集上的BLEU值大于0.25。

二、网络实现

网络结构使用pytorch官网上的tutorial的seq2seq+attention结构。

encoder用GRU模型, decoder用attention模型。

1、项目结构

- data文件夹:包含训练集、测试集、验证集的数据,和最后得到的模型。
- image文件夹:包含得到的图像和报告里的图像。
- config文件夹:包含训练网络的一些主要参数。
- BLEU.py文件: 负责计算BLEU值
- main.py文件: 负责运行整个网络的训练
- test.py文件:负责将读取测试集、验证集的数据,将xml格式的数据转为txt文本格式的数据并——对应。

其中运行main.py文件时需要指定config的路径。

其中config参数包括:

```
"test-de-path": "./data/de_test.txt",
  "test-en-path": "./data/en_test.txt",
  "dev-de-path": "./data/de_dev.txt",
  "dev-en-path": "./data/en_dev.txt",
  "train-de-path": "./data/news-commentary-v11.de-en.de",
  "train-en-path": "./data/news-commentary-v11.de-en.en",
  "MAX-LENGTH": 20,
  "train": true,
  "teacher-forcing-ratio": 0.5
}
```

2、数据预处理

期初在数据预处理阶段就遇到了一点问题,样本数据量很大,而且样本中的英文文本中夹杂一些印地语,导致德文和英文的行数不是——对应的,这个坑导致—开始取全部数据的训练过程很失败。

```
mews-commentary-v11.de-en.en - i已事本

文件(F) 編集(E) 稽式(O) 查看(V) 帮助(H)

The current selection procedure is losing legitimacy in a changing world, and it carries gre ^
The consequences of retaining it might not be as dramatic as Keynes' prophesied, but the समलैंगिक पुरुषों को जोखिम-पूर्व रोग-निवारण (PrEP) के लिए तैयार करना लंदन — अक्तूबर में, एचआईवी के खिलाफ लड़ाई में एक संभावित महत्वपूर्ण खोजवाली दवा के प्रभाव पर शोध करनेवाले शशीधकर्ताओं ने यह पाया कि जो समलैंगिक पुरुष पुरुषों से यौन संबंध बनाते समय कंडोम का उपयोग करने के अलावा दूर इस प्रकार एचआईवी/एड्स के प्रसार को रोकने के लिए लड़ रहे लोगों के शस्तागार में यह एक नया उपकरण आ गया है। इस साल की गर्मियों में, विश्व स्वास्थ्य संगठन ने इस उद्देश्य से एक महत्वपूर्ण कदम उठाया, उसने सभी समलैंगिक पुरुषों र लेकिन इसमें आनेवाली एक महत्वपूर्ण बाधाः विकासशील दुनिया के अधिकतर भाग में समलैंगिक पुरुषों के कानूनी स्थिति नाइजीरिया में आधिकारिक रूप से स्वीकृत समलैंगिकता के विरोध के वातावरण ने एड्स के खिलाफ लड़ाई में पहले ही अ परिणाम इससे अधिक गंभीर नहीं हो सकते थे। दो साल पहले, एक युवा एचआईवी पॉजिटिव नाइजीरियाई पुरुष ने मुझे क्लं मेरा फेसबुक का दोस्त ही ऐसा अकेला व्यक्ति नहीं है जिसे नाइजीरिया के समलैंगिकता के विरोध की कीमत चुकानी पड़ी एचआईवी के साथ जीनेवाले नाइजीरियाइयों को केवल संक्रमण से लड़ने के बजाय कुछ और भी करना चाहिए, उन्हें सामां
```

通过比对发现数据集中的前两万条数据是可以使用的,于是便用前两万条数据进行训练。

使用编码建立英文、德文词典,并映射为one-hot编码。

```
class Lang:
    def __init__(self, name):
        self.name = name
        self.word2index = {}
        self.word2count = {}
        self.index2word = {0: "SOS", 1: "EOS"}
        self.n_words = 2 # Count SOS and EOS
```

```
def addSentence(self, sentence):
    for word in sentence.split(' '):
        self.addWord(word)

def addWord(self, word):
    if word not in self.word2index:
        self.word2index[word] = self.n_words
        self.word2count[word] = 1
        self.index2word[self.n_words] = word
        self.n_words += 1
    else:
        self.word2count[word] += 1
```

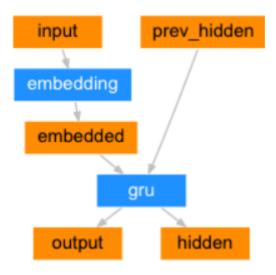
采用One-Hot编码的原因是简单快捷,效果也不差。

3、网络搭建

使用GRU循环神经网络,一开始想用LSTM网络,但LSTM网络参数太多,训练太慢,在个人电脑上训练时间允许。还考虑过使用Transform模型,但又不是很了解Transform模型,忙于复习考研也没有太多时间钻研学习Transform模型,于是就使用了pytorch官网上的tutorial模型,并加了一些小小的修改。

ENCODER

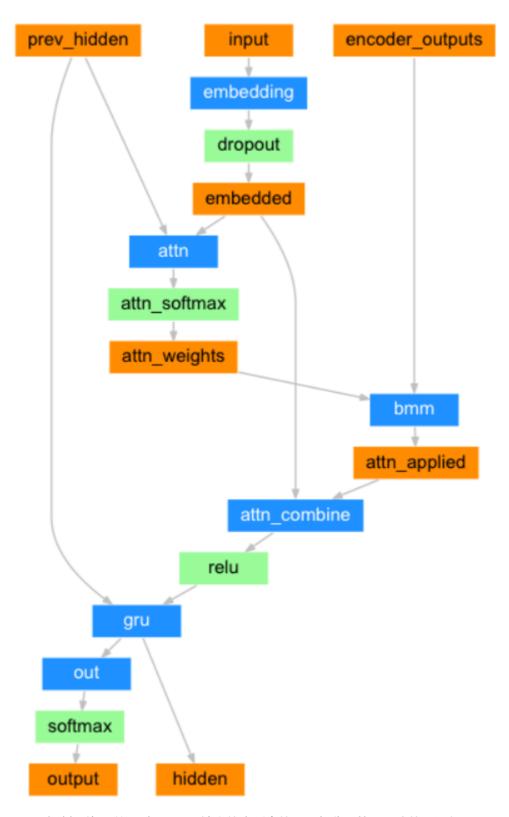
encoder采用传统的GRU模型,大体上和tutorial上的示例一样。



embedding层为我们之前的训练好的one-hot词向量,prev_hidden隐层设置hidden_size为256.

DECODER

decoder采用使用attention机制的GRU模型,和tutorial上的示例大体一致。



通过attention机制网络可以识别encoder输出的序列中的不同部分,从而正确的翻译句子。

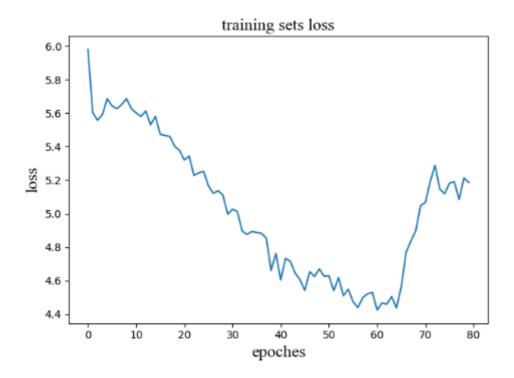
3、训练网络与实验结果

对于循环神经网络来说,训练网络是最痛苦的事情,还好我的笔记本上有一颗英伟达1050ti的GPU,训练速度会快很多。

• 训练环境:

СРИ	GPU	内存	CUDA
Intel Corei7 2.20GHz	NVIDA GeForce GTX 1050TI(4G显存)	16GB	11.0

- 训练时间:2h,
- **训练数据量**:由于训练数据集的对其问题,其2w行之前是对齐的,但2w行之后数据就对不齐了, 所以我们只取训练数据集的前两万行进行训练。
- 训练次数: 每轮epoch进行1000轮训练, 总共进行80000轮训练即80轮epoch。
- 训练loss:



可以看到在训练集上的loss先下降,差不多在60次的时候loss达到最小值,之后loss反而上升,所以我们取60次epoch时得到的模型。

• 训练结果:在训练集上有的句子BLEU值为0.3几,有的句子的BLEU值却很低

```
74m 24s (- 0m 0s) (80000 100%) 3.7201
> timing is also an issue especially with respect to saving incentives .
= auch der zeitpunkt ist ein problem insbesondere in bezug auf sparanreize .
< ein problem ist ein problem mit dem problem mit problem problem .<EOS>
r: 12
BLEU值为, 0.30769230769230765
> we need a democratic competitive environment initiative at all levels an active civil society and real public control .
= wir brauchen ein demokratisches wettbewerbsorientiertes umfeld initiative auf allen ebenen eine aktive zivilgesellschaft und echte offentliche kontrolle .
< wir brauchen eine eine eine eine eine und und eine und und . . <EOS>
r: 18
BLEU值为, 0.27578028205768607
> they successfully portrayed themselves as victims of a firestorm rather than as accessories to arson .
= sie haben sich erfolgreich als opfer eines flachenbrands dargestellt anstatt als gehilfen bei der brandstiftung .
< sie sie sie ein sie eines als eines von als als als als . <EOS>
r: 16
BLEU值为, 0.3118356616772059
> what should the new government s economic policy agenda be ?
= wie soll die wirtschaftspolitische agenda der neuen regierung aussehen ?
```

每条句子的BLEU值不是很稳定,有的句子BLEU值很高,有的句子BLEU值很低,而且对BLEU值取平均会发现BLEU值低的句子占比更大。

紧接着在测试集上进行测试,将config文件中的train参数改为false,在运行main.py文件

```
> one of his two friends ran to him and begged the attacker to let the victim go .

= als seine beiden freunde herbeieilten und beschwichtigend auf den schlager einredeten lie dieser zunachst von seinem opfer ab .

< seine anderen dem mit dem anderen dem . zu . . . . . <EOS>
r: 19

BLEU值为: 0.10458938416503243

> he accused the european nations of collectively looking the other way .

= er warf den europaischen staaten kollektives wegschauen vor .

< er gibt aber mit anderen anderen europaischen . <EOS>
r: 9

BLEU值为: 0.3333333333333333333333
```

可以看到在测试集上bleu值也是忽高忽低很不稳定。平均一下同样可以发现bleu值偏低的句子占大多数。

其原因我觉得可能是因为我只取了前2w条数据进行训练,训练的数据量不够,而且只训练了8w轮,训练数量和强度不够,导致模型没有充分学到数据中的内容。

但这数据集对不齐,只能用前两万行,而且用的是个人电脑,虽然有一个已经过时了的GPU但训练速度还是太慢了,更多的训练数据和训练量需要的训练时间更是无法忍受的。所以就不拿更多的数据进行训练了,得到的在测试集上的平均bleu值为0.31

L: 1/

BLEU平均值为: 0.3144966240478535

三、实验总结

本次实验具体实践了GRU循环网络的实现,和它在seq2seq和机器翻译上的应用,虽然大部分代码是借鉴pytorch上的tutorial的,但结合课上的理论知识,也能明白其中的道理,照着tutorial上的代码理解一遍,更加深了课上所学的理论只是。这次实验让我认识到硬件速度对深度学习这类算法的重要性,即使数据量和参数的训练量不是很大的情况下,在个人电脑也要跑相当长一段时间。这就导致个人学习的障碍,但Google推出的colab等免费GPU平台稍微降低了个人学习的成本,但要开vip离线训练还是要收费滴,而且还要挂梯子。