

The background features a light blue gradient with abstract geometric patterns. On the left and right sides, there are complex structures made of overlapping triangles in shades of beige, brown, and grey, connected by black lines. Scattered throughout the background are numerous thin, light grey triangles of various sizes and orientations.

小牛寒假实习汇报

■ 汇报人：印飞

低精度相关工作

一

CPU环境下编写Float16

二

相关文献阅读

三

分类器替换理论部分

四

分类器替换实验部分

Pre-work

Samaritan-Infi / LowPrecision

Watch 1

Star 1

Fork 0

Code

Issues 0

Pull requests 0

Projects 0

Wiki

Insights

Settings

Branch: master

LowPrecision / LowPrecision / src /

Create new file

Upload files

Find file

History

Samaritan-Infi Create FNNLM.h

Latest commit 7063c8b just now

FNNLM.cpp

Create FNNLM.cpp

a minute ago

FNNLM.h

Create FNNLM.h

just now

Float16_with_calculation.cpp

Create Float16_with_calculation.cpp

2 months ago

HalffloatToFloat.cpp

Create HalffloatToFloat.cpp

2 months ago

README.md

Create README.md

2 months ago

README.md

通过编码实现float16的类型 float16为半精度类型，和CUDA库中标准一样 1个标志位，5位阶码，10位尾码

- CPU: Float16
- 与Float32的转换 & 计算

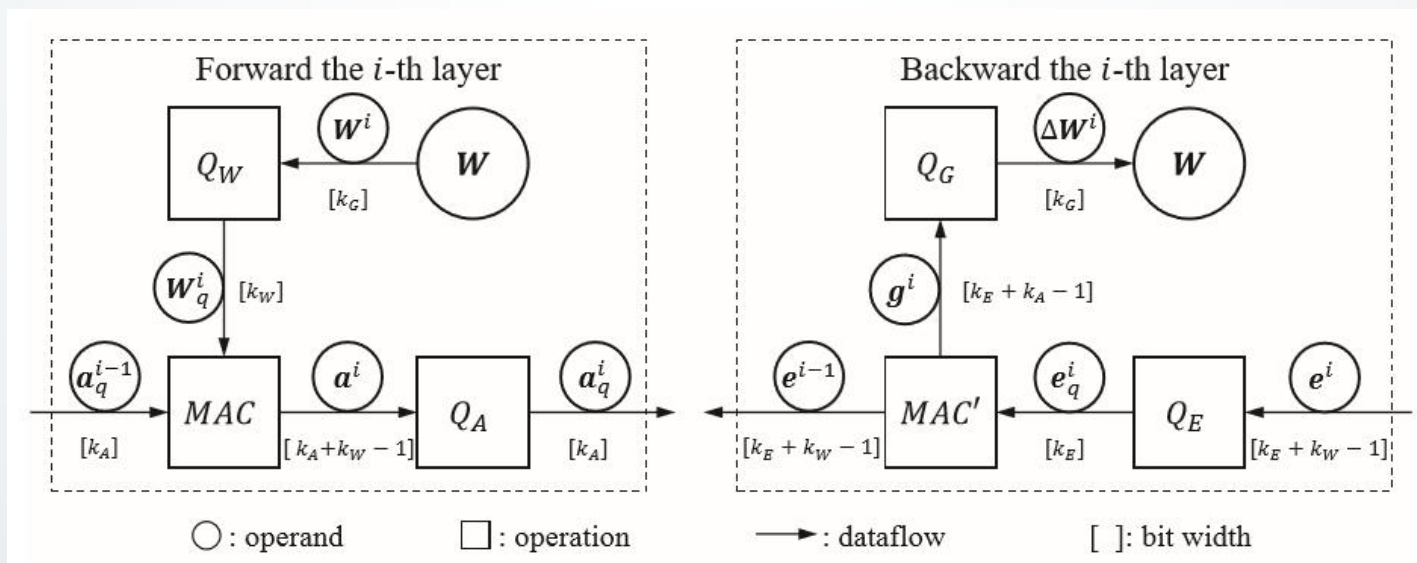
半精度类型，参照CUDA库中定义标准
1个标志位，5位阶码，10位尾码



Introduction

- 精度降低至float16 / int8
- 二值网络
- 参数修剪

WAGE



Related Work

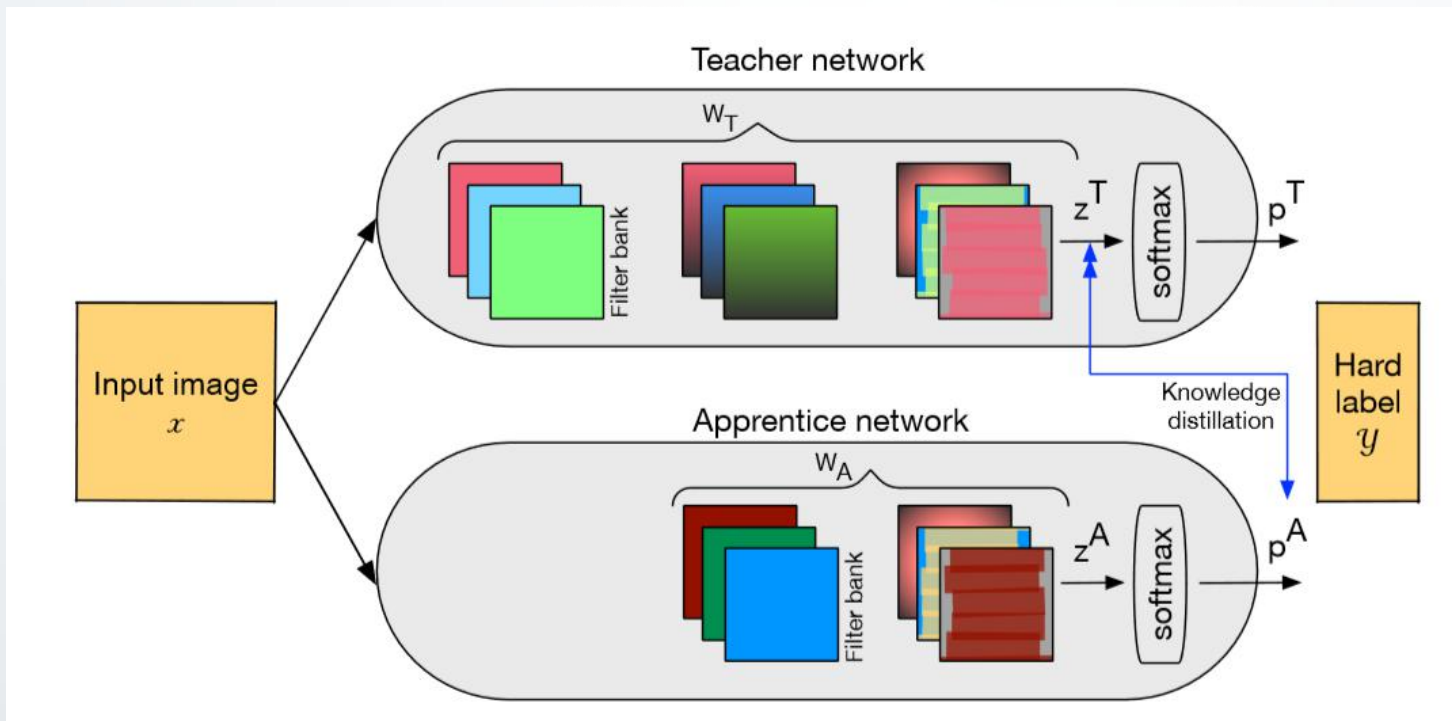
知识蒸馏 (knowledge distillation)

将知识从一个复杂的网络转移到一个更小的网络中。

low-precision

$$\mathcal{L}(x; W_T, W_A) = \alpha \mathcal{H}(y, p^T) + \beta \mathcal{H}(y, p^A) + \gamma \mathcal{H}(z^T, p^A)$$

$\mathcal{H}(\cdot)$ 表示损失函数



Related Work

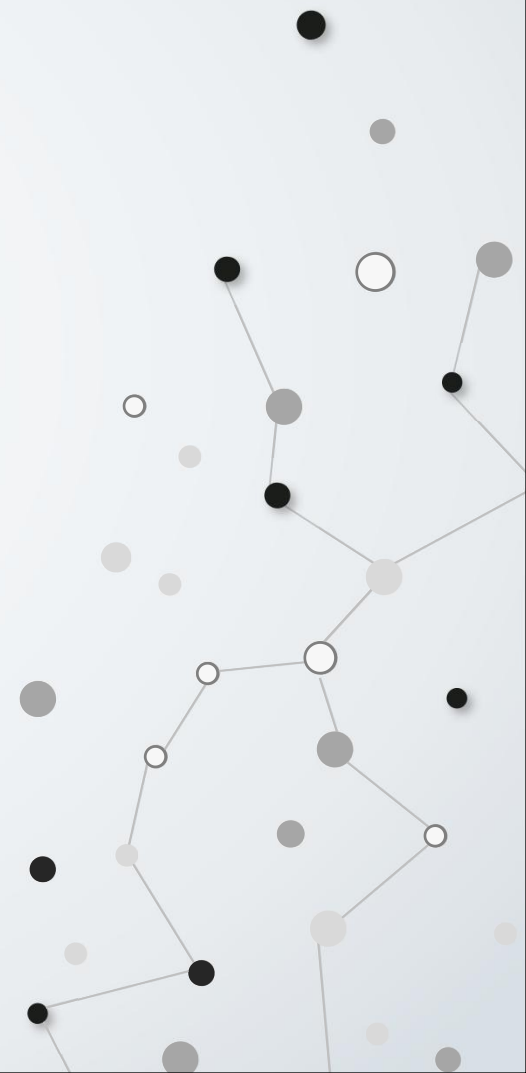
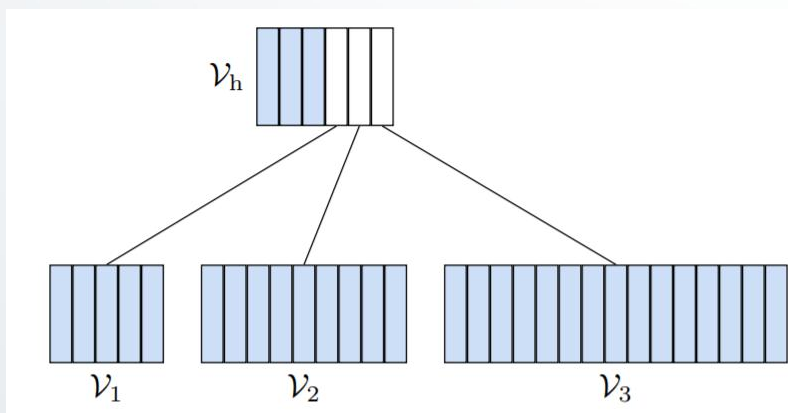
层次 softmax (hierarchical softmax)

$$p(w | h) = p_1(C(w) | h) \times p_2(w | C(w), h)$$

复杂度 $O(dk)$ 降低至 $O(d\sqrt{k})$

适应性 softmax (adaptive softmax)

通常情况下，20% 的词表可以覆盖 87% 的文档内容



目标：使整形的精度能够通过分类器，以提高速度

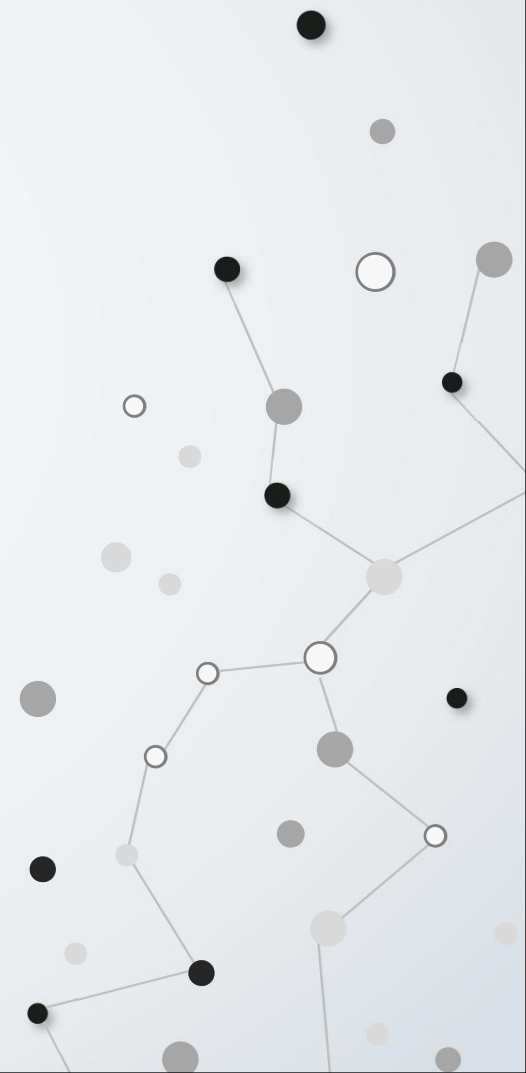
$$\log_softmax(z_w) = \log\left(\frac{\exp(z_w)}{\sum_{w' \in V} \exp(z_{w'})}\right)$$

初始，尝试直接通过线性函数直接去拟合分类公式

包括泰勒展开， k^x

$$e^x = \sum_{n=1}^{\infty} \frac{x^n}{n!} \quad \ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n$$

由于泰勒拟合的函数只在一定定义域内保持性质，超出之后会发生突变，在语言模型中测试时无法收敛。



Theory

$$\log_softmax(z_w) = \log\left(\frac{\exp(z_w)}{\sum_{w' \in V} \exp(z_{w'})}\right)$$

$$Float_w \xrightarrow{scale(x)} Int_w$$

$$Log_softMax(z_w) = \log(\exp(z_w)) - \log\left(\sum_{w' \in V} \exp(z_{w'})\right)$$

$$Log_softMax(z_w) \approx z_w - \log\left(\sum_{w' \in V} \exp(z_{w'})\right)$$

$$\sum_{w' \in V} 2^{z_{w'}} = a_0 2^0 + a_1 2^1 + \dots + a_n 2^n (a_i \in \{0,1\}, a_n = 1)$$

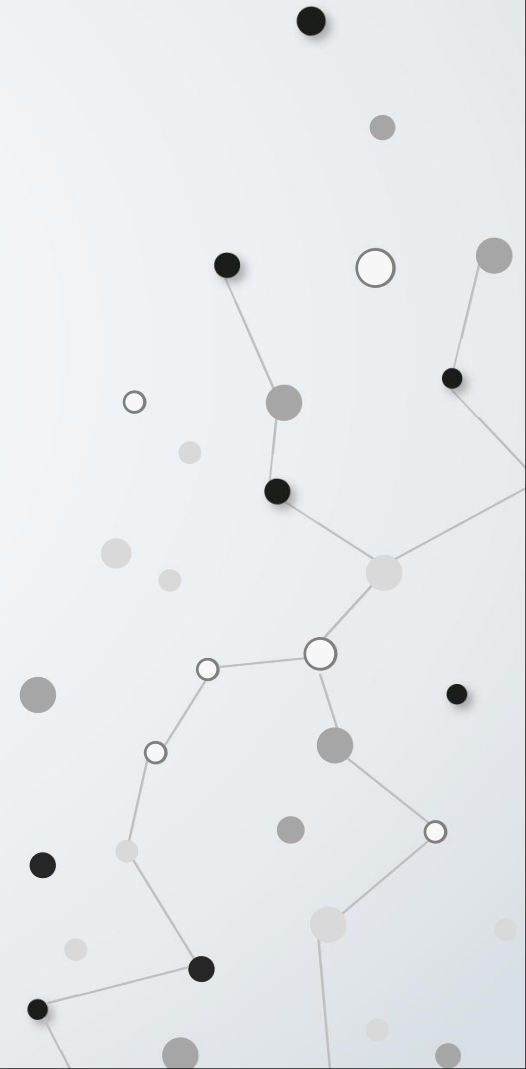
$$\sum_{w' \in V} \exp(z_{w'}) \approx a_0 \exp(0) + a_1 \exp(1) + \dots + a_n \exp(n) (a_i \in \{0,1\}, a_n = 1)$$

$$\exp(n) \leq \sum_{w' \in V} \exp(z_{w'}) \leq \exp(n+1)$$

$$n \leq \log\left(\sum_{w' \in V} \exp(z_{w'})\right) \leq n+1$$

$$n = \max(z_{w'}) + b, b \geq 0$$

$$Log_softMax(z_w) \approx z_w - \max(z_{w'}) - b, b \geq 0$$



Experiment

在Neu.Trans中FNNLM的语言模型上尝试修改

修改了前向过程最后输出的分类器

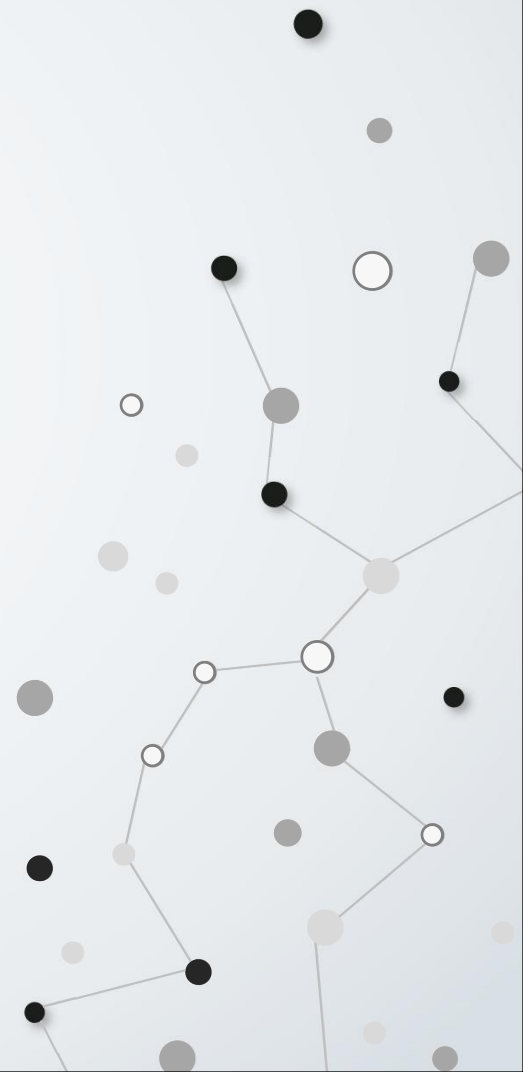
收敛时间较基线明显缩短

统计计算时间

在一个epoch内:

基线计算softmax的时间为 1.723192 秒

尝试修改后的计算时间为 0.956978 秒



Experiment

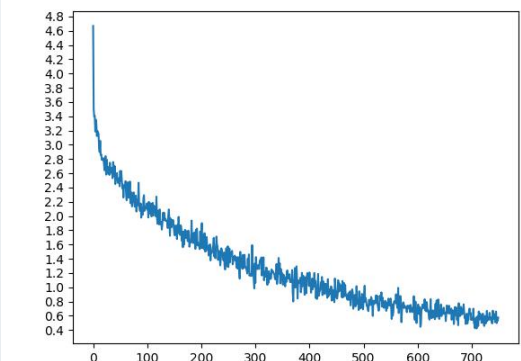
在一个基于RNN和Attention机制 Sequence to Sequence 的翻译模型上修改

修改了decoder中输出的分类器

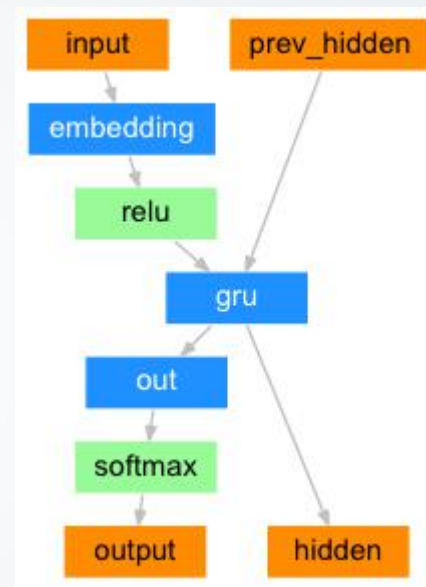
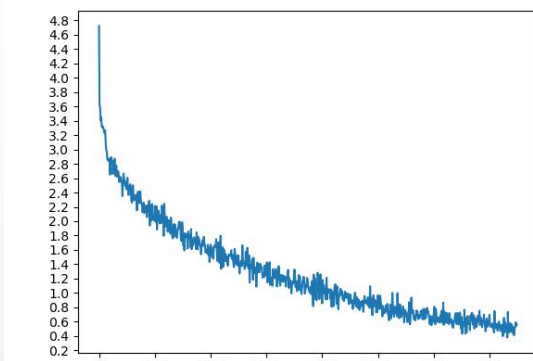
decoder的输出会影响到训练的反向过程

loss图像如下， 前后收敛过程相近

初始



修改后



Experiment

在一个基于RNN和Attention机制 Sequence to Sequence 的翻译模型上修改

训练数据量 10599

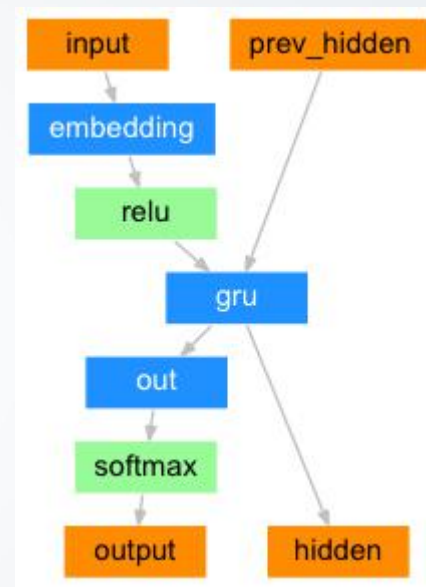
测试数据 1000

(数据均为简单句子)

测试BLUE值结果:

初始: 72.83

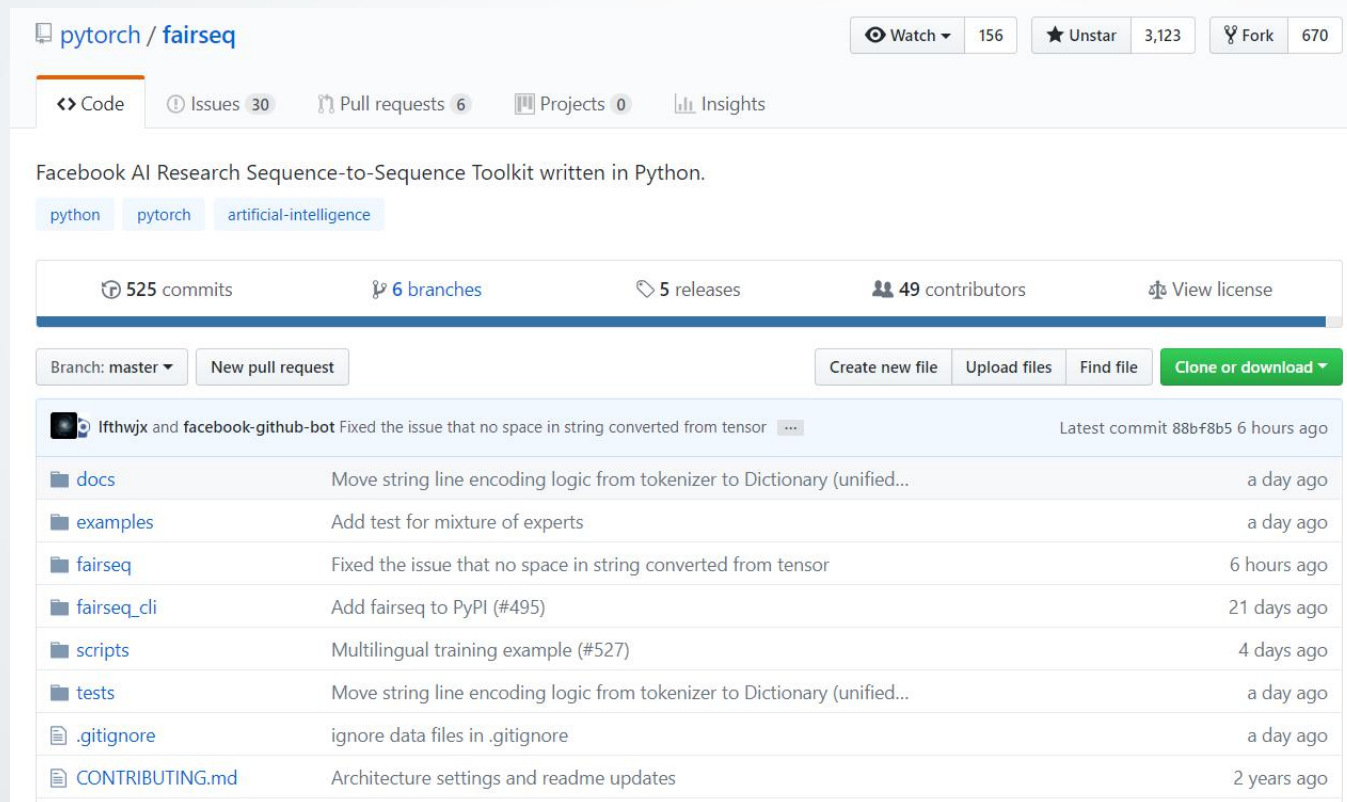
修改后: 73.04



Experiment

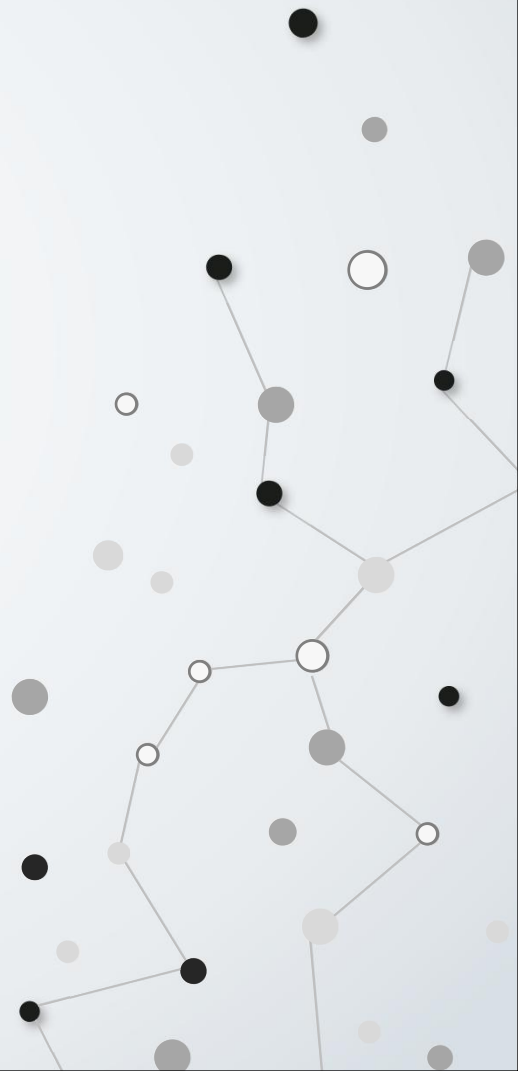
在FaceBook的fairseq翻译模型上尝试修改

这里直接采用它提供的预先训练的模型进行测试，直接修改其最后的decoder输出部分的分类器



The screenshot shows the GitHub repository page for `pytorch/fairseq`. At the top, it displays the repository name and statistics: 156 watchers, 3,123 stars, and 670 forks. Below this, there are tabs for Code, Issues (30), Pull requests (6), Projects (0), and Insights. The description states: "Facebook AI Research Sequence-to-Sequence Toolkit written in Python." and includes tags for `python`, `pytorch`, and `artificial-intelligence`. A progress bar shows 525 commits, 6 branches, 5 releases, and 49 contributors. Below the progress bar, there are buttons for "Branch: master", "New pull request", "Create new file", "Upload files", "Find file", and "Clone or download". The commit history table shows the latest commit by `lfthwjx` and `facebook-github-bot` fixing an issue with string conversion, followed by a list of recent commits and their descriptions.

Commit	Message	Time
88bf8b5	Fixed the issue that no space in string converted from tensor	6 hours ago
	Move string line encoding logic from tokenizer to Dictionary (unified...	a day ago
	Add test for mixture of experts	a day ago
	Fixed the issue that no space in string converted from tensor	6 hours ago
	Add fairseq to PyPI (#495)	21 days ago
	Multilingual training example (#527)	4 days ago
	Move string line encoding logic from tokenizer to Dictionary (unified...	a day ago
	ignore data files in .gitignore	a day ago
	Architecture settings and readme updates	2 years ago



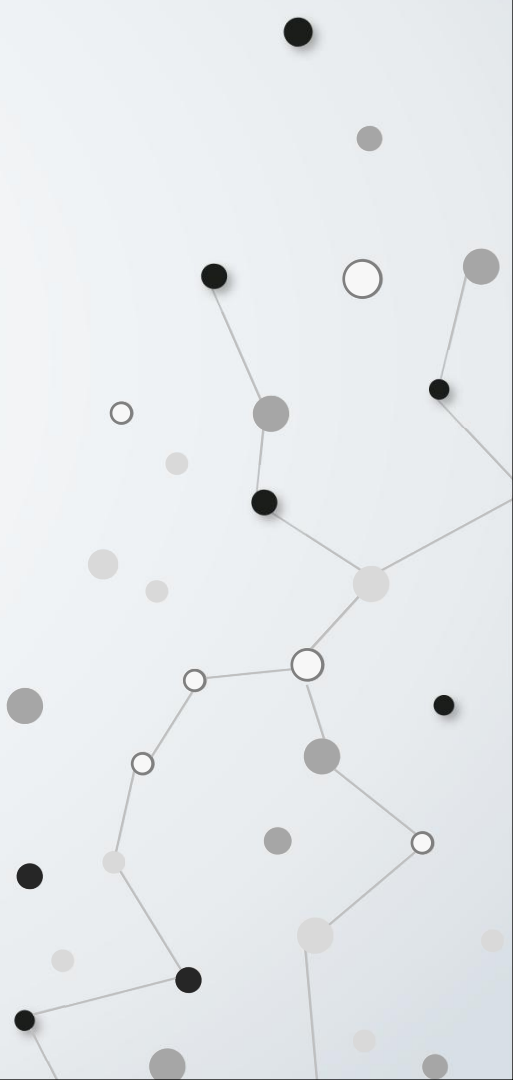
Experiment

在FaceBook的fairseq翻译模型上尝试修改

这里直接采用它提供的预先训练的模型进行测试，直接修改其最后的decoder输出部分的分类器

Pre-trained models 测试模型

Description	Dataset	Model	Test set(s)
Convolutional (Gehring et al., 2017)	WMT14 English-French	download (.tar.bz2)	newstest2014: download (.tar.bz2) newstest2012/2013: download (.tar.bz2)
Convolutional (Gehring et al., 2017)	WMT14 English-German	download (.tar.bz2)	newstest2014: download (.tar.bz2)
Transformer (Ott et al., 2018)	WMT14 English-French	download (.tar.bz2)	newstest2014 (shared vocab): download (.tar.bz2)
Transformer (Ott et al., 2018)	WMT16 English-German	download (.tar.bz2)	newstest2014 (shared vocab): download (.tar.bz2)



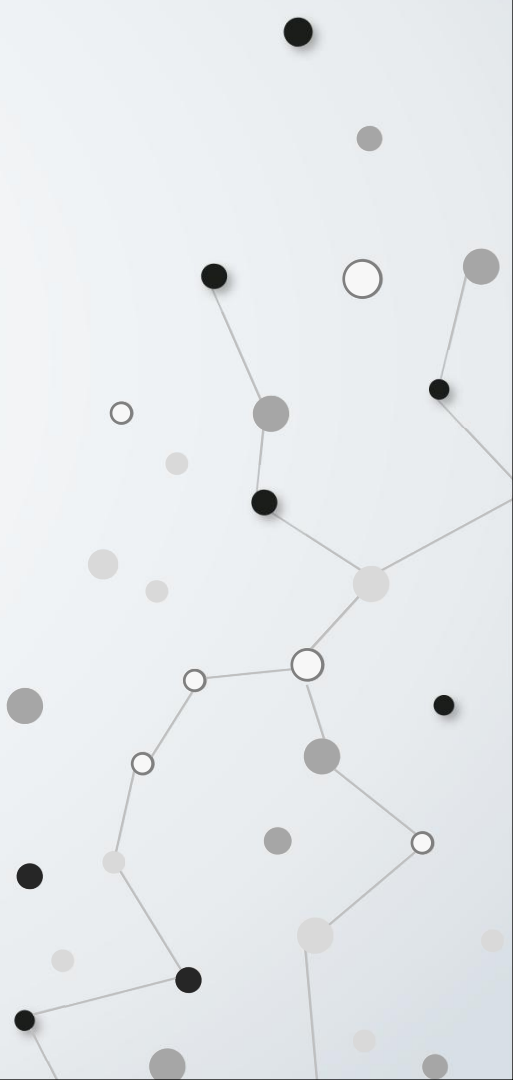
Experiment

在FaceBook的fairseq翻译模型上尝试修改

这里直接采用它提供的预先训练的模型进行测试，直接修改其最后的decoder输出部分的分类器

测试结果

Model	Origin BLUES	Now BLUES	Percent
CNN - WMT14 English-French	40.83	40.08	98.163%
CNN - WMT14 English-German	25.70	24.80	96.498%
Transformer - WMT14 English-French	43.00	42.26	98.279%
Transformer - WMT16 English-German	29.23	28.68	98.118%



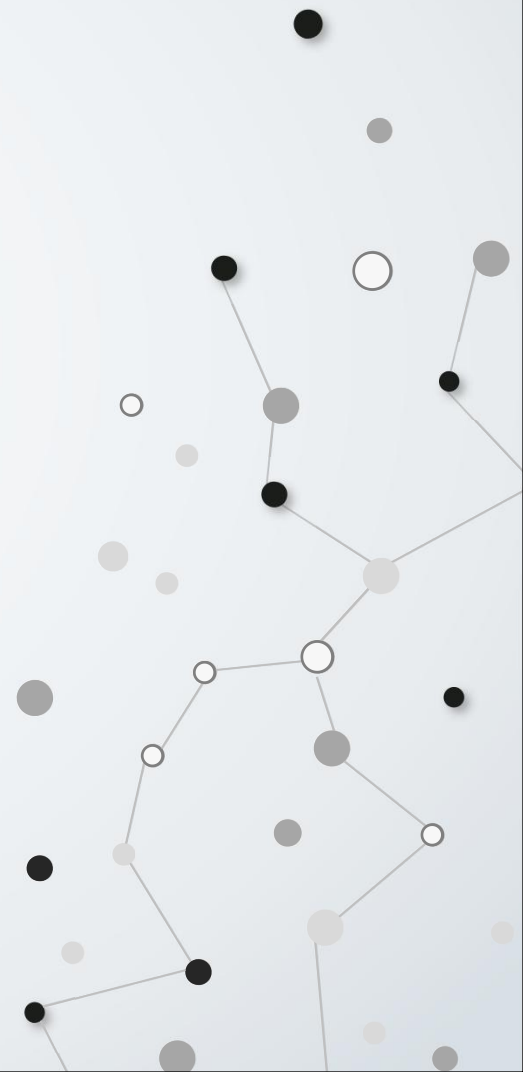
Conclusion



为了能够支持整形精度的解码，采用了线性方式的分类器来替代原先的softmax。

替换后，支持整数精度，时间上也有所下降，但是准确性目前效果并不理想，后续应有更多的学习，尝试与完善。

当指数分类器能够被较为理想的替换后，可以尝试将整个decoder替换为整形精度，测试其速度与准确率。



感谢聆听



AIM AT