

革命性提升-宇宙最强的NLP预训练BERT模型（附官方代码）

忆臻 机器学习初学者 2018-12-05



编辑 忆臻

公众号 | 机器学习算法与自然语言处理 yizhennotes

1. Bert官方源码公开

终于是在千呼万唤始出来，Google AI 发表于10月中旬的论文：

《BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding》一下子在NLP领域击其千层浪。文中提出的BERT模型，在11项NLP任务（包括阅读理解，文本分类、推断，命名实体识别等）中都取得了start of art 的突破性成绩！

这个成绩着实吓死了一批研究人员，其中的一些任务也可以说宣布没有什么研究空间了。

截止发稿前，短短时间，BERT已经获得近8k star，可见其受关注程度。

google-research / bert

Watch 571 Star 8,803 Fork 1,550

Code Issues 50 Pull requests 7 Projects 0 Wiki Insights

TensorFlow code and pre-trained models for BERT <https://arxiv.org/abs/1810.04805>

nlp google natural-language-processing natural-language-understanding tensorflow

84 commits 1 branch 0 releases 19 contributors Apache-2.0

Branch: master New pull request Create new file Upload files Find file Clone or download

jacobdevlin-google Merge pull request #170 from imcaspar/patch-1 Latest commit 4a47cc2 7 days ago

.gitignore	Initial BERT release	a month ago
CONTRIBUTING.md	Initial BERT release	a month ago

2. 项目仓库包含的内容

- 用于BERT模型架构的TensorFlow代码（主要是标准的Transformer架构）。
- BERT-Base和BERT-Large模型小写和Cased版本的预训练检查点。
- 论文里微调试验的TensorFlow代码，比如SQuAD，MultiNLI和MRPC。
此项目库中的所有代码都可以直接用在CPU，GPU和云TPU上。

jacobdevlin-google Fix SQuAD hyperparams and improve README		Latest commit a08ff75 5 hours ago
.gitignore	Initial BERT release	16 hours ago
CONTRIBUTING.md	Initial BERT release	16 hours ago
LICENSE	Initial BERT release	16 hours ago
README.md	Fix SQuAD hyperparams and improve README	5 hours ago
__init__.py	Initial BERT release	16 hours ago
create_pretraining_data.py	tree-wide: minor typo fixes	15 hours ago
extract_features.py	tree-wide: minor typo fixes	15 hours ago
modeling.py	Updating documentation and adding requirements.txt	9 hours ago
modeling_test.py	Initial BERT release	16 hours ago
optimization.py	Initial BERT release	16 hours ago
optimization_test.py	Initial BERT release	16 hours ago
requirements.txt	Updating requirements.txt to make it only 1.11.0	9 hours ago
run_classifier.py	tree-wide: minor typo fixes	15 hours ago
run_pretraining.py	tree-wide: minor typo fixes	15 hours ago
run_squad.py	Updating run_squad.py to reduce CPU memory usage	6 hours ago
sample_text.txt	Initial BERT release	16 hours ago
tokenization.py	Initial BERT release	16 hours ago
tokenization_test.py	Initial BERT release	16 hours ago

3. 大家关心的问题，是否支持其它语言（如汉语）

目前放出的预训练模型是英语的，我们大家肯定都会关心是否会有汉语或者其它语言预训练model的公布。

多语言模型支持的语言是维基百科上语料最大的前100种语言（泰语除外）。多语言模型也包含中文（和英文），但如果你的微调数据仅限中文，那么中文模型可能会产生更好的结果。

就是这里列出的1-60号语言：

https://meta.wikimedia.org/wiki/List_of_Wikipedias#All_Wikipedias_ordered_by_number_of_articles

4. 最后再看看BERT的屠榜和官方代码地址

Introduction

BERT, or Bidirectional Encoder Representations from Transformers, is a new method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing (NLP) tasks.

Our academic paper which describes BERT in detail and provides full results on a number of tasks can be found here: <https://arxiv.org/abs/1810.04805>.

To give a few numbers, here are the results on the [SQuAD v1.1](#) question answering task:

SQuAD v1.1 Leaderboard (Oct 8th 2018)	Test EM	Test F1
1st Place Ensemble - BERT	87.4	93.2
2nd Place Ensemble - nlnet	86.0	91.7
1st Place Single Model - BERT	85.1	91.8
2nd Place Single Model - nlnet	83.5	90.1

And several natural language inference tasks:

System	MultiNLI	Question NLI	SWAG
BERT	86.7	91.1	86.3
OpenAI GPT (Prev. SOTA)	82.2	88.1	75.0

Plus many other tasks.

Moreover, these results were all obtained with almost no task-specific neural network architecture design.

If you already know what BERT is and you just want to get started, you can [download the pre-trained models](#) and [run a state-of-the-art fine-tuning](#) in only a few minutes.

地址点击：<https://github.com/google-research/bert>

论文 (<https://arxiv.org/abs/1810.04805>)

作者公众号：



长按二维码扫描关注

机器学习算法与自然语言处理

ID: yizhenotes

通俗笔记， 分享交流

请关注和分享↓↓↓



机器学习初学者

QQ群: 774999266或者654173748 (二选一)

往期**精彩**回顾



- 机器学习简易入门-附推荐学习资料
- 机器学习初学者公众号下载资源汇总 (一)
- 黄海广博士的github镜像下载 (机器学习及深度学习资源)
- 吴恩达老师的机器学习和深度学习课程笔记打印版
- 机器学习小抄- (像背托福单词一样理解机器学习)
- 首发: 深度学习入门宝典-《python深度学习》原文代码中文注释版及电子书
- 科研工作者的神器-zotero论文管理工具
- 机器学习的数学基础
- 机器学习必备宝典-《统计学习方法》的python代码实现、电子书及课件