Light Pre-Training Chinese Language Model for NLP Task

CLUE Benchmark

Mar. 2020

With the development of Deep Learning, pre-training models plays an increasingly important role in CV and NLP community. Since Google released BERT in 2018, pre-training method became normally in NLP researches. A series of pre-training models based on Transformer has been emerged and occupied a critical position. It was not hard to find that those models were large along with great order of magnitude. On the one hand, they required lots of computational and memory resources in training phase and service stage separately. On the other hand, inference speed of them cannot reach the desired height. These two demerits limited their utilization in industry.

Academic and industrial community have been considering how to solve these problems effectively. Some modified the structure of neural network. For example, ALBERT reduces parameters by factorized embedding parametrization and Cross-layer parameter sharing. Distillation is also a good idea. The Knowledge Distillation means that the teacher model with deep networks and great number of parameters teach the smaller one, which is called student here, so as to compress the knowledge into the student model, obtaining good results with the reduction of model size. There are certainly many other works, which put forward different ideas from different angles, and finally get good results under the condition of reducing model size.

1. Task setting

The main purpose of this task is to train a powerful model with limited model size to get the best possible result in all four types of tasks provided by us.

This task covers four different downstream tasks, including sentence-pair classification, Coreference Resolution, Named-entity recognition, and reading comprehension. For details about datasets, please visit the website: https://github.com/CLUEbenchmark/LightLM#dataset-description

【Chinese Corpus】：

To meet the challenge of the lack of Chinese Corpus, we provide 14G data for participants.

https://github.com/CLUEbenchmark/CLUECorpus2020#cluecorpussmall14g

2. Register：

- Visit www.CLUEbenchmark.com, and click the button 【注册】 at the top right corner of the page. After that, please log in.
- After selecting the【NLPCC 测评】in the top navigation bar, please register our task in 【比赛注册】.
- If you have any questions, please contact us via： CLUEBenchmark@163.com .

3. Submission and metrics

Submission

- Visit www.CLUEbenchmark.com.
- Click 【立即测评】, then submit your results.
- You could check your grade on 【NLPCC2020 小模型】 in 【排行榜】

Submission example could be found on https://storage.googleapis.com/cluebenchmark/tasks/nlpcc_task1_submit_examples.zip.

Metrics

Preliminary rounds:

Model size：＜12M。（1/9 * 110M）。

Inference time：1/8 * (time of bert-base) on the same environment

Under the setting above, the average scores on all four tasks provided by us will be the only one criterion.

Finals:

We will reproduce the results of each team in the finals, and they will be given grades according to the formula below:

$$\mathbf{res} = \mathbf{P/100 * 0.8 + 0.1 * \left(1 - \frac{S_{Lite}}{0.9 * S_{\mathit{Bert}}}\right) + 0.1 * \left(1 - \frac{T_{Lite}}{T_{Bert}}\right)}$$

P: scores on average

S: Size of model

T: Average time for inference given all the test data

_ bert: bert base

_Lite: light model

【Note】:

The name of your submission should contain “nlpcc”, e.g. nlpcc-ELECTRA.

4. Timeline

Warmup: 3/25 - 4/5

Participants could warm them up on others tasks on the website such as https://www.cluebenchmarks.com/small_model_classification.html. This could be a great chance to become familiar with this system.

Preliminary rounds: 4/6-5/15

We will release the final tasks before 4/6:

- CLUEWSC2020（The Winograd Schema Challenge,Chinese Version）

- CSL (Keyword Recognition from paper)
- CLUENER2020 （NER）
- CMRC 2018(Reading Comprehension)

People could submit their results to this system, and check their grades on the ranking list.

Ranking List：https://www.cluebenchmarks.com/nlpcc2020.html

Finals： 5/15-5/20： Same tasks as Preliminary rounds

Evaluation： 5/20-5/30

We will reproduce the results of some teams in the finals. Participants have to prepare the shell which could be easily run to reproduce the results on our machines.

Result: The results will be published with other shared tasks in NLPCC.

5.Reward

- 1st：RMB 10, 000， 1 team
- 2nd：RMB 5000， 1 team
- 3rd：RMB 2500， 1 teams

The top 3 participating teams of this task will also be certificated by NLPCC and CCF Technical Committee on Chinese Information Technology.

6.Related sites

NLPCC2020 official site:
http://tcci.ccf.org.cn/conference/2020/cfpt.php

CLUEbenchmark

www.cluebenchmark.com

github

https://github.com/CLUEbenchmark/LightLM

7.Others

nlpcc-light模型task交流榜

该二维码7天内(4月1日前)有效，重新进入将更新

Reference :

[1] https://github.com/CLUEbenchmark/DistilBert

[2] Turc, Iulia, et al. "Well-read students learn better: The impact of student initialization on knowledge distillation." arXiv preprint arXiv:1908.08962 (2019).

[3] Yang, Ziqing, et al. "TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing." arXiv preprint arXiv:2002.12620 (2020).

[4] Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." arXiv preprint arXiv:1909.11942 (2019).

[5] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[6] Clark, Kevin, et al. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." International Conference on Learning Representations. 2019.

[7] Jiao, Xiaoqi, et al. "Tinybert: Distilling bert for natural language understanding." arXiv preprint arXiv:1909.10351 (2019).

[8] Cheng, Yu, et al. "A survey of model compression and acceleration for deep neural networks." arXiv preprint arXiv:1710.09282 (2017).

[9] Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya. "Reformer: The Efficient Transformer." arXiv preprint arXiv:2001.04451 (2020).

[10] https://mrqa.github.io/

[11] http://tcci.ccf.org.cn/conference/2019/taskdata.php

[12] https://mp.weixin.qq.com/s/HdG3_CaSdZP3lCp8J_VRQA

[13] https://digital.csic.es/bitstream/10261/163973/1/Performance_Analysis_of_Real_Time_DNN_on_RPi.pdf

[14] https://ibug.doc.ic.ac.uk/resources/lightweight-face-recognition-challenge-workshop/

NLPCC-2020 轻量级中文语言模型评测

CLUE Benchmark

2020 年 3 月

自深度学习流行以来，预训练模型在图像或者自然语言处理领域的作用变得越来越大。自然语言处理领域的的预训练主要是自Bert诞生之后流行了起来。基于Transformer的各种预训练模型在各大NLP相关任务中逐渐成为主流。但是，相应的，这些模型很大，参数量很多，所以不仅训练起来需要很大的资源，在部署使用的时候也需要很多的资源，最直观的就是内存的使用。同时，在追求性能的时候，推理时间一直不够快。这两个缺陷一直在制约此类模型在实际生产环境中的使用。

当然，学术界和工业界一直在思考如何才能一定程度上解决这个问题。有人从网络模型入手，例如Albert通过矩阵分解，和权值跨层共享减少参数；蒸馏模型利用知识丰富的教师教相对小的学生模型，从而将知识压缩到较小的模型中，在减少模型参数的情况下获取好的效果。当然还有很多其他的工作，从不同的角度提出了不同的思路，最终都在减少模型规模的情况下拿到了不错的结果。

一、任务设置

本次任务的主要目的是在限定参数数量的情况下，尽可能的训练一个性能足够好的模型，能够在我们提供的四个类型的任务中都得到比较好的效果。

本次任务覆盖了四种不同的下游任务，包含 sentence-pair 分类，指代消解，命名实体识别和阅读理解。限于篇幅，具体的数据集介绍，请登陆网站查看介绍以及数据示例。

https://github.com/CLUEbenchmark/LightLM#dataset-description

【中文数据集】：

为了缓解参赛者对中文语言的需求，我们提供了 14G 的中文语料供大家选用。

https://github.com/CLUEbenchmark/CLUECorpus2020#cluecorpussmall14g

二、任务注册：

1. 访问 www.CLUEbenchmark.com， 右上角点击【注册】并登录。
2. 进入【NLPCC2020】tab 页，选中【注册栏】后进行比赛注册，并点击提交。
3. 如果有注册或者各个方面有关比赛的问题，可以通过官方邮箱：CLUEBenchmark@163.com 联系我们。

三、结果提交和评价

1. 访问 www.CLUEbenchmark.com，点击立即评测，提交结果。
2. 在排行榜中选择 NLPCC-小模型榜可以查看自己的排名。
3. 提交自己的模型计算结果，提交示例请查看：https://storage.googleapis.com/cluebenchmark/tasks/nlpcc_task1_submit_examples.zip

评估标准：

- 预赛评估：

参数量限制：小于 12M。（1/9 * 110M）。

推理速度：在同样设备条件下，8 倍于 bert-base 的推理速度。

在满足以上两个限制条件下，使用在我们提供的四个任务上的平均得分作为评估依据。

- 决赛评估：

我们会根据参赛情况，对最终决赛的队伍的进行代码复现，并根据以下公式对计算最终得分。

$$\mathbf{res} = \mathbf{P/100 * 0.8 + 0.1 * \left(1 - \frac{S_{Lite}}{0.9 * S_{\mathit{Bert}}}\right) + 0.1 * \left(1 - \frac{T_{Lite}}{T_{Bert}}\right)}$$

P: scores on average

S: Size of model

T: Average time for inference given all the test data

_ bert: bert base

_Lite: light model

【提交要求-重要】：提交的结果名称应该包含”nlpcc”，例如：nlpcc-ELECTRA。

四、比赛时间设置

Warmup：3/25 - 4/5

参赛者可以在 https://www.cluebenchmarks.com/small_model_classification.html 上进行尝试提交，熟悉系统并测试自己的模型的效果

预赛：4/6-5/15

我们会在这个 4/6 之前释放最新的任务组合:

1. CLUEWSC2020（Winograd 模式挑战中文版）
2. CSL 论文关键词识别（Keyword Recognition)
3. CLUENER2020（命名实体识别）
4. CMRC 2018(阅读理解)

参赛者可以提交自己的结果到比赛榜单，在榜单上看到自己的排名。

比赛榜单：https://www.cluebenchmarks.com/nlpcc2020.html

决赛：5/15-5/20：决赛采用和预赛相同的任务

评估：5/20-5/30

我们会根据参赛情况，要求最终决赛的队伍的提交模型与运行脚本并由我们进行结果复现，并计算最终得分。要求参赛者提供可以在 linux 服务器上直接运行的脚本，并指明依赖库，以及对应的硬件依赖。

结果公布：结果会与 NLPCC 的其他 shared task 结果一起公布。

五、奖励

第一名：奖励人民币：1 万元整，共一名
第二名：奖励人民币：5 千元整，共一名
第三名：奖励人民币：2.5 千元整，共一名
前三名队伍会获得 NLPCC 和 CCF 中国信息技术技术委员会认证的证书。

六、相关网站

NLPCC2020 官方网站
http://tcci.ccf.org.cn/conference/2020/cfpt.php
CLUEbenchmark 官方网站
www.cluebenchmark.com
本次比赛 github
https://github.com/CLUEbenchmark/LightLM

七、其他

比赛交流群：

nlpcc-light模型task交流榜

该二维码7天内(4月1日前)有效，重新进入将更新

参考文献：

[1] https://github.com/CLUEbenchmark/DistilBert

[2] Turc, Iulia, et al. "Well-read students learn better: The impact of student initialization on knowledge distillation." arXiv preprint arXiv:1908.08962 (2019).

[3] Yang, Ziqing, et al. "TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing." arXiv preprint arXiv:2002.12620 (2020).

[4] Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." arXiv preprint arXiv:1909.11942 (2019).

[5] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[6] Clark, Kevin, et al. "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." International Conference on Learning Representations. 2019.

[7] Jiao, Xiaoqi, et al. "Tinybert: Distilling bert for natural language understanding." arXiv preprint arXiv:1909.10351 (2019).

[8] Cheng, Yu, et al. "A survey of model compression and acceleration for deep neural networks." arXiv preprint arXiv:1710.09282 (2017).

[9] Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya. "Reformer: The Efficient Transformer." arXiv preprint arXiv:2001.04451 (2020).

[10] https://mrqa.github.io/

[11] http://tcci.ccf.org.cn/conference/2019/taskdata.php

[12] https://mp.weixin.qq.com/s/HdG3_CaSdZP3lCp8J_VRQA

[13] https://digital.csic.es/bitstream/10261/163973/1/Performance_Analysis_of_Real_Time_DNN_on_RPi.pdf

[14] https://ibug.doc.ic.ac.uk/resources/lightweight-face-recognition-challenge-workshop/