谢撩，人在斯坦福打SoTA

原创 Jazon 夕小瑶的卖萌屋 2021-05-28 17:00

文 | Jazon
编 | 小戏

> 小编注：不知道大家还记不记得卖萌屋之前人在斯坦福，刚上CS224n的Jazon小哥发来的关于斯坦福神课CS224n上半学期的报道？今天，Jazon又在斯坦福前线发来了关于他在CS224n下半学期的经历，那么现在让我们把画面交给Jazon，看看大佬的课程作业是怎么完成的吧！

上篇文章提到我在 **Stanford** 上 **NLP “神课” CS224n**，课程的前半学期以上课、写作业为主，而后半学期则基本都是 **Guest Lectures** ，没有作业，让我们专心做 **Project** 。

往年，时不时都有 **224n** 的 **Project** 转化为**顶会 Paper** ，**224n** 学期结束的海报展览也是 **Stanford** 校园里的“**学术盛典**”之一。今年可惜展览取消了，不过大家的 **Reports** 仍然都会放在官网上。能让我的 **Report** 发表在 **Stanford** 课程的官网上，这是一种很大的荣幸呀！**所以我积攒了足够的动力，一定想尽力做好这次 Project** 。

Project 规则

自从 2016 年斯坦福创造的问答数据集 **SQuAD** 横空出世，几年来 **224n** 的 **Project** 都是分为 “**Default**” （默认）、“**Custom**”（自定义）两种选项，其中 **default** 就是搞 **SQuAD** ，**Custom** 则没什么选题的限制。

2021 年的 **default Project**，分成了 2 个 **Track**：一个叫 **IID**（就是单纯地做 **SQuAD**）；还有一个叫 **RobustQA**，**这个更有意思一些，需要学生搭建的 SQuAD 模型有“鲁棒性”**，比如给定另一个数据集 **NewsQA** 极少量的训练数据，能够在 **NewsQA** 上也取得好的结果。

IID 不可以用任何基于 **BERT** 的模型，**RobustQA** 也只能用 **DistilBERT**，两者都不能使用给定训练集以外的数据。这很容易理解，不这样的话，**Project** 很容易变得没有挑战性。**但这些限制也让我感觉，两个 Default Project 不够刺激，于是，我就选择了做 Custom Project。**

Project 可以一个人做，也可以组成 2 ~ 3 人的小队。今年课上的 477 个学生一共分成了 275 个小组（不少人都是 Solo 做），每个组有一定的 **Azure 云计算 Credits**。我的队友 **Nina** 是一位（很稀有的）来自中国的 **Stanford** 本科生，现在读大四。

选题

确定做 **Custom Project** 以后，选题并不是件容易的事。**从宏观上说，需要知道哪些事情需要做、可以做；要了解之前的人做过什么、没做过什么。这些都需要投入不少精力。**

对于 **Custom Project**，除了自己选题以外，还有一个列表，表上列出了 34 个课题的 **Proposal**，分别由近 20 位 **Stanford** 在读 **PhD** 提出，做列表上的课题就由对应的 **PhD** 指导。这些 **Projects** 最诱惑人的地方在于，有些导师是明确希望把 **Project** 转化为长期合作的项目与顶会 **Paper** 的。

列表里第 20 个课题叫“**Can Transformers Do Math**”，我觉得有些意思，就联系了那位 **PhD**，问他现在还接不接受指导新的组。但后来觉得这个题目不够 NLP，而且感觉这位 **PhD** 同时指导很多组会很忙，所以就放弃了。

我们最终决定做的 **Project** 还是来自小轶的提议。1 月底，我问小轶对 **224n Project** 有何建议，她推荐了 2 个方向，其中一个是**多人对话 QA**，比如有个数据集叫 **FriendsQA** [1]，是一个基于老友记台词构建的 **QA** 数据集。我觉得这个 **Topic** 非常有意思，和队友讨论之后，就决定做它啦（PS：2 月 8 号下午，我们在 **Manning** 的 **Office Hour** 上问他这个方向如何，不过他并没有探索过这个领域，就没有给出什么实质性的建议）。

决定做 **FriendsQA**，我知道这其中的风险：**Manning 在课上讲过，有时没有选好题的 Custom Project 会做得比较“Lame”**。确实，**FriendsQA** 是一个很新的数据集，整个这个小

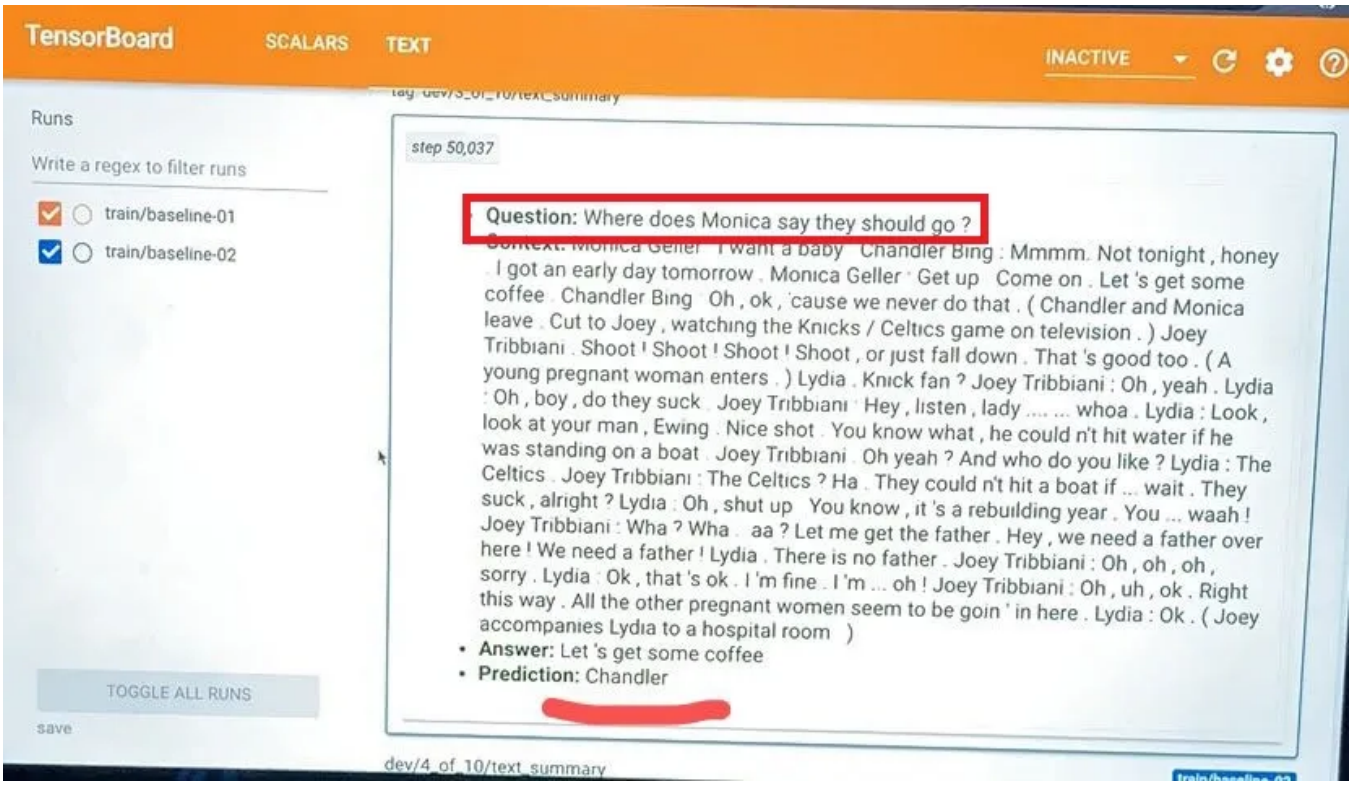
领域也很少有人进行研究，这让我对于 **Project** 能做成什么样完全没有把握。**但正是因为没啥人做过，做起来才刺激嘛**！我在日记里写到：

> “见证、体验了太多搞科研受到各种条件的限制后，现在我在 224n 的平台上，能自由地探索世界上还没有人能很好解决的难题，就像 Google Brain、MSR、FAIR 的那些大牛一样享受真正的学术自由，这是难得的浪漫和奢侈。”

开启 Project

2 月 15 号提交了 **Proposal**，之后搞完前文提到的让全班爆炸的作业 5，休息了几天，2 月底正式开干！

一开始，我们是完全按照 **SQuAD** 的思路来做的，把 **FriendsQA** 转化成和 **SQuAD** 一样的格式以后，跑各种用在 **SQuAD** 上的模型。**但奇怪的是，每次一开始训练，在 FriendsQA 上的 F1 分数就断崖式下跌**。那段时间，看模型胡说八道的 **Output** 一度成为我和队友的欢乐源泉。



小轶的判断是代码出了 **Bug**，到了 3 月 6 号我才发现确实如此，**是我在把 FriendsQA 格式转化成 SQuAD 时，标准答案的 Start_index 算错了**！我以为是按 **Word** 的位置算，原来是按 **Char** 的位置算。**这个 Bug 并不好找，因为跑 Evaluation 时计算准确率并没有问题（算准确率只看词是否匹配，不看词的位置）**。

寻找方向

解决掉 **Bug** 以后，我们的模型取得了不错的 **F1** 分数。但显然，只跑 **BERT** 对于满足 **Final Project** 的要求是远远不够的，我们得做一些技术含量更高的改进。

Project Handout 里面的建议是：**仔细钻研一个方法好于粗浅地尝试很多方法**。可是怎么找到要钻研的那个方向，又是个令人头大的问题。

一开始，我和 **Nina** 主要研究了怎么改 **BERT** 的结构（**具体来说，在 BERT 里加入 Utterance-Level Embedding，ULE**），但最后得出的结论是，我们不太可能发明什么其他的模型结构或者训练方法，可以做得好过那篇研究 **FriendsQA** 的 ACL paper，《**Transformers to Learn Hierarchical Contexts in Multiparty Dialogue for Span-based Question Answering**》[2]了。

改 **BERT** 行不通，那该怎么办？我想了很多，可以做的思路实在太多：**用 BERT 以外的模型？从数据入手？研究更好的评估指标？一时间，我和队友都有些迷茫焦虑。**

我在需要动力的时候，会去油管上看 **Carykh** 的视频：他分别把他在 **Stanford** 两门 AI 课上做 **Project** 的经历做成了视频。其中在 **CS230**（深度学习）课上，他的小组几经周折，直到 **Project** 的最后一周才确定了选题 [3]。

pyq 里也经常能看到一起搞 **224n** 的同学吐槽“**机器拒绝学习**”，还有一次一起出去买东西，开车的同学在等一个红灯时想着他的 **Project**，突然灵光一闪，就恍然大悟地拍了一下方向盘，然后在手机上匆忙记了点笔记。**现在想想，真是好有趣**。

确定方向

话说我们的任务和 **Default** 选项的 **RobustQA Track** 十分相似，都需要根据较少的训练数据，搭建 **Robust** 的 **QA** 系统。于是，我们开始从 **RobustQA** 的 **Handout** 上寻找灵感。里面提供了一些可探索的方向，**如“ Mixture-of-Experts ”、“ Domain Adversarial Training ”、“ Meta Learning ”**。

当时我有搞 **RobustQA** 的同学，已经试过了这些方法，但很多都使效果不增反降，只有 **Data Augmentation**（数据增强）看上去有前途。于是3月7号，我和 **Nina** 开了个会，终于决定了我们的最终计划：接下来一周专攻数据增强，如果搞好了，最后一周可以冲刺用 **Ensemble** 刷分。

此时，离 Report 截止还有不到12天，我便开启了猛肝 Project 的模式。

创新的快感

按照我俩的分工，我主攻的数据增强方法是 **Back-Translation** （反向翻译）。**它的原理很简单，就是一段英语句子，翻译成比如说汉语，再回翻成英语，就得到了一段意思一致、表述不同的新句子。**

对于像 **SQuAD** 、 **FriendsQA** 这样的 **QA** 数据集来说，每道问题的标准答案，都是从给定的某段上下文里截取的一个 **Span** 。问题来了，现在上下文的用词变了，这个 **Span** 怎么重新找呢?

这个问题似乎探索过的人不多，在 **QANet** 那篇 **Paper** 《**QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension**》[4]里用的方法是比较“ **Character-Level 2-Gram Score** ”，虽然效果不错，但显然它只是个 **Heuristics** ——只看字母的匹配，而没有考虑含义是否匹配。有没有更好的办法呢?

3 月 10 号，我一直想着这个问题，突然有了个 idea ： **为啥不把所有词向量相加起来，做比较和匹配呢**? 这样不仅考虑进了（词组/句子）的意义，也考虑了 **Span** 的长度。这个简单的新方法似乎并没有人用过，于是我初步跑了一些实验，实验效果还不错。这让我十分高兴，这是我首次体验到创新的快感(下图就是我们最后提交的 **Summary Diagram** ，是这个新方法的示意图)!

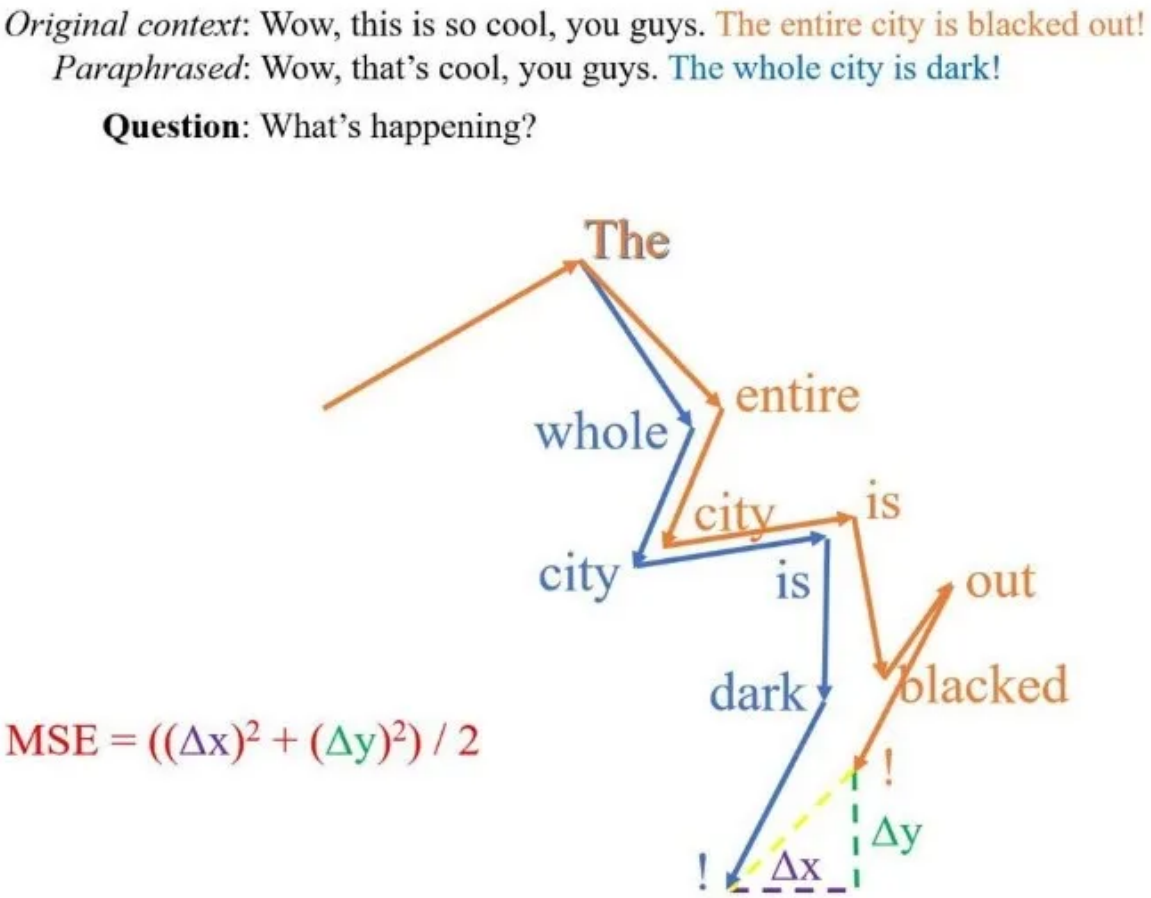
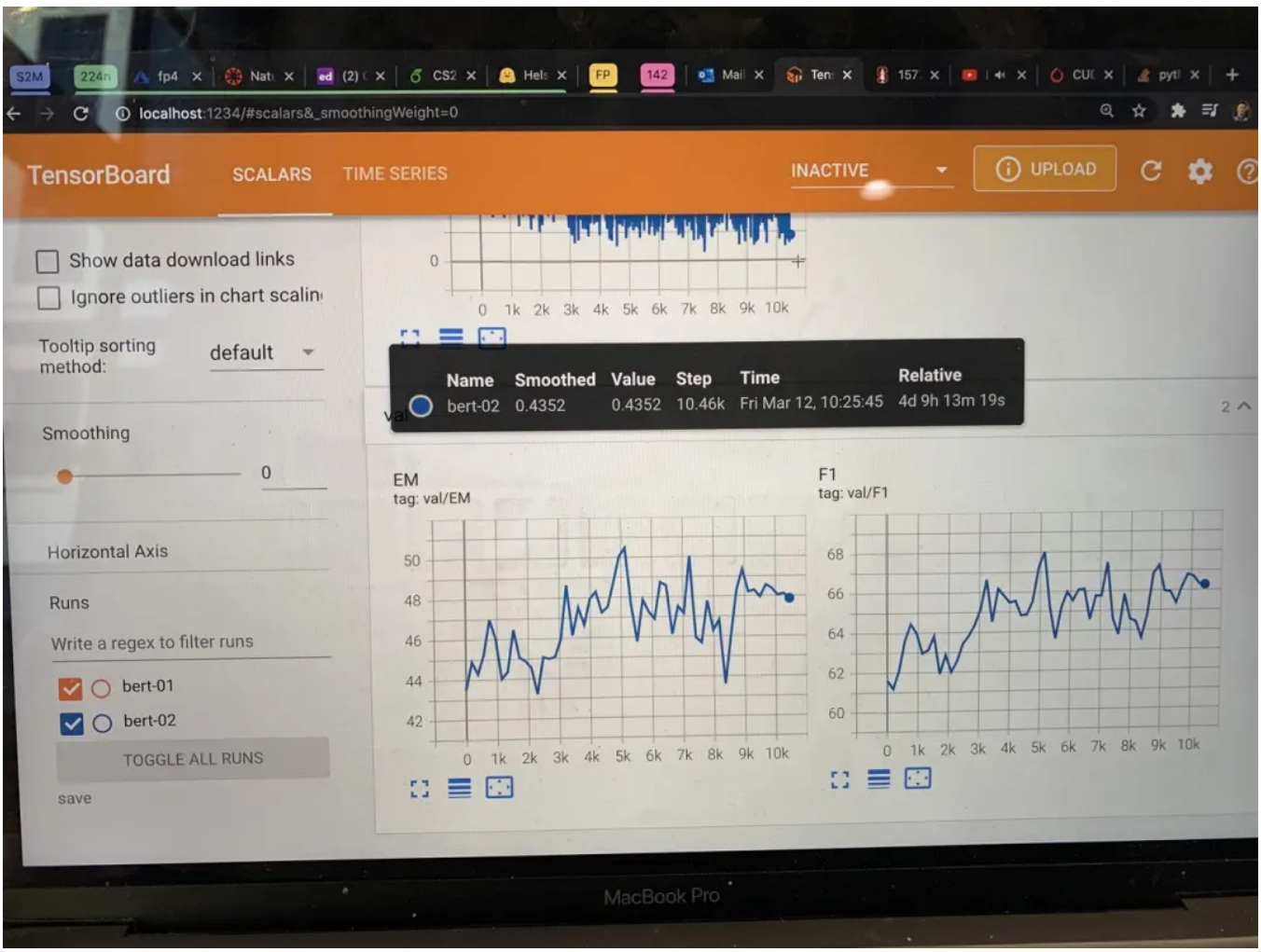


Figure 1: 2D illustration of our sum-of-word-vectors technique for finding the answer in a paraphrased context, used for data augmentation. Orange represents the original answer, and blue indicates the paraphrased answer that we wish to find. The span with the smallest MSE is selected as the paraphrased answer.

击败SoTA

生成了几个增强数据集、验证新方法的可行性之后，最后一周，按照计划，我们用上 **BERT-Large** 和 **Ensemble** ，冲刺 **FriendsQA** 的 **SoTA** 。

那时我就像炒股一样，天天盯着 Loss 和 F1 ，看是涨了还是跌了... Project做到这里，已经快做吐了，不过看着离SoTA越来越近，还是很激动。



最终，我们的 **F1** 分数达到了 72.1 ，比原 **SoTA** 提高了 2.5 个百分点。诚然，我们用了比原 **SoTA** 模型多好几倍的参数，才达到这些成绩；**但对于一次 Final Project 来说，时间、资源如此有限，我觉得已经很不容易啦。**

3 月 19 号凌晨，我和 **Nina** 肝完了 **Report**，一共 5000 多词写满了 8 页纸。这里我特别想提的是，几乎所有的顶会 **Paper**，都会因为篇幅限制，省略超级多的细节，这可能会使得复现比较麻烦。

而在课程的 **Project Report** 里，我们可以补全这些细节，事无巨细地描述实验是怎么做的：比如，计算词向量时，如果出现 **OOV** 词汇怎么处理？**Ensemble** 具体如何综合不同模型的 **Predictions**？**Train** 、**Dev** 、 **Test** 三个 **Set** 具体怎么用？等等。

总之，我在这次 **NLP Project** 上前后一共投入了大约 140 小时的精力，2021 年 3 月算是我两年多以来最忙碌的一段日子，**每天最多能在 Project 上投入 15 小时**，当时还有其他课要上、有 **Lab** 的工作要做。但做完以后真的成就感满满！

这段经历，可以说第一次让我全方位地体验了科研的乐趣。这里也感谢小轶提供的指点，我觉得比我们课程分配的 **Mentor** 有用多啦（当然这不怪我们 **Mentor**，毕竟他得一个人指导十几个 **Team**...）

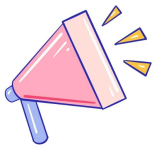
评奖出炉

224n 结课一个月之后，4 月 17 号，TA 们整理好了所有的 **Report**、选出了获奖的 **Report**，都公开在了 **224n** 的官网上。我们届 **MSCS** 有两位女生做的医学图像转文本的 **Custom Project**，拿到了最佳报告奖，非常厉害！

我们组的 **Report**[5] 可以在这里找到：

> **报告题目：**
> **Data Augmentation and Ensembling for FriendsQA**
>
> **报告链接：**
> *https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/reports/final_reports/report072.pdf*

224n 之后的春学期有一门 **224u（NLU）**，算是 **224n** 这门课的延续，不过我因为排课的原因并没有上，比较可惜。**不管怎样，我在 224 的体验非常棒，希望以后可以给大家带来更多在 Stanford 计算机系上课的见闻！**



后台回复关键词【**入群**】

加入卖萌屋NLP/IR/Rec与求职讨论群

后台回复关键词【**顶会**】

获取ACL、CIKM等各大顶会论文集！

参考文献

[1] FriendsQA: Open-Domain Question Answering on TV Show Transcripts.
https://www.aclweb.org/anthology/W19-5923.pdf

[2] Transformers to Learn Hierarchical Contexts in Multiparty Dialogue for Span-based Question Answering
https://arxiv.org/pdf/2004.03561.pdf

[3] AI Lip Reading:
https://youtu.be/28U6EwfKois

[4] QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension:
https://arxiv.org/abs/1804.09541

[5] Data Augmentation and Ensembling for FriendsQA:
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1214/reports/final_reports/report072.pdf

Allen AI提出MERLOT，视频理解领域新SOTA！

夕小瑶的卖萌屋