NBME – Score Clinical Patient Notes

一、研究背景与方向

1.1 研究背景

本次研究基于**本小组参加的**机器学习网站Kaggle的比赛，比赛题目为NBME - Score Clinical Patient Notes（给临床病历打分），副标题为Identify Key Phrases in Patient Notes from Medical Licensing Exams（从执业资格考试中识别患者笔记中的关键短语）。下面是该竞赛的背景描述：

当患者去看医生时，他们如何解释您的症状可以决定您的诊断是否准确。到他们获得许可时，医生已经进行了大量练习，编写患者记录，记录患者的投诉历史、体检结果、可能的诊断和后续护理。学习和评估写病人笔记的技能需要其他医生的反馈，这是一个耗时的过程，可以通过结合机器学习来改进。

直到最近，“第二步临床技能考试”成为**美国医学执照考试（USMLE）**的一个组成部分。该考试要求应试者与标准化病人（受过训练的人，以描绘特定的临床病例）互动，并写下病人的笔记。训练有素的医生评分员随后用概述每个病例的重要概念（被称为特征）的评分标准对病人笔记进行评分。在病历中发现的这种特征越多，分数就越高（除其他因素外，还包括对考试的最终得分的贡献）。

然而，让医生对患者笔记考试进行评分需要大量时间以及人力和财力资源。已经创建了使用**自然语言处理**（NLP）的方法来解决这个问题，但患者笔记的计算评分仍然具有挑战性，因为特征可能以多种方式表达。例如，特征“对活动失去兴趣”可以表示为“不再打网球”。其他挑战包括需要通过组合多个文本片段来映射概念，或者存在与关键基本要素如“缺乏其他甲状腺症状”相对应的模棱两可的否定词，例如“没有感冒不耐受、脱发、心悸或震颤”。

本次研究的目标是开发一种自动识别每个患者笔记中相关特征的方法，特别关注从标准化患者访谈中捕获含特征的信息。

1.2 研究方向

经过对题目的分析，本小组认为可以从一个方向切入，即可以认为该研究是一种QA（Question Answer）问题，即给定患者笔记和特征，设计算法，提取笔记中反应特征的段落。目前针对QA问题，常用的方法有两种：第一种是预测“答案”段落的开始（Start）和结束（End）位置，第二种是对原文中每个词（tocken）进行二分类，判断其是否为“答案”段落的一部分。

根据现实问题，本小组发现，有时需要通过组合多个文本片段来映射概念，即反应特征的“答案”可能是几个不连续的文本片段，所以我们认为对原文中每个词进行二分类的方法更为合适。

目前随着大规模预训练模型如transformer，bert，GPT等应用到NLP领域，用预训练模型作为初始参数并进行**微调（fine-tuning）**的方式大大减小了问题解决的成本，因此本文也拟基于上述研究范式展开研究。

二、相关技术

注意力机制（Attention）由Bengio团队于2014年提出并广泛应用；2017年，Google团队基于Attention机制提出了可训练神经网络**Transformer**，该神经网络仅由Attention及其相关的变体层以及前馈神经网络（Feed Forward Neural Network）堆叠组成，该网络创新了Seq2Seq结构，并在机器翻译的BLEU中取得了SOTA效果；基于Transformer的编码层，Google在2018年提出了**Bidirectional Encoder Representation from Transformers(BERT)**，该模型进一步堆叠了Transformer的编码层，克服了GPT模型”“masked“的缺陷，并使用了基于Books和Wikipedia的大规模语料，使得模型参数超过100M，在13个下游任务中取得了SOTA效果；同时，

论文还提出了基于BERT的fine-tuning方法，大大降低了基于下游任务建模的难度，实现了将研究基线从单纯使用词嵌入（word embedding）到使用词嵌入和预训练模型参数的转变；2021年，微软基于BERT提出了**Decoding-enhanced BERT with Disentangled Attention（DeBERTa）**，改善 BERT 和 RoBERTa预训练效率，并提升了下游指标，在SuperGLUE这项自然语言理解基准任务上“超越人类”，以90.3分夺冠。

另外，BERT在发展过程中出现了诸如RoBerta，ALBert，PubMed（医学领域）。

三、模型介绍

本文的研究将主要使用**DeBERTa-V3和PubMed**进行。下面是主要技术的解释：

3.1 DeBERTa

该算法在BERT的基础上，提出了两项创新点。

1. **自注意力解耦机制(disentangled self-attention mechanism)**。即每个word的embedding 由content embedding 和 和 position embedding 组成，但不是BERT的content embedding 和 position embedding 直接向量相加。 word之间的注意力权重用word的content和相对位置的解耦矩阵表示。
2. 用**增强的Masked Decoder**，替换原始输出的softmax层，用来预测模型预训练时设置的掩码词（masked word），以此解决BERT预训练和微调阶段不一致的问题。

首先来看自注意力解耦机制，给定文本序列，对第i个$tocken$分别用H_i和$P_{i|j}$表示文本向量和相对位置j位置的位置向量，$tocken\ i,j$之间的注意力权重（cross attention score）可以分解为：

$$
\begin{aligned}
A_{i,j} &= \{H_i, P_{i|j}\} \times \{H_j, P_{j|i}\}^T \\
&= H_i H_j^T + H_i P_{j|i}^T + P_{i|j} H_j^T + P_{i|j} P_{j|i}^T
\end{aligned}
$$

解耦的最终结果是4种注意力机制，即content2content，content2position，position2content，position2position。由于使用的是相对位置编码，position2position这一项并不能提供更多额外的信息，所以实现过程将其忽略。

以单头为例，标准的自注意力表示如下：

$$
Q = HW_q, K = HW_k, V = HW_v, A = \frac{QK^T}{\sqrt{d}}
$$

$$
H_0 = softmax(A)V
$$

当最大相对位置为k，那么tocken i,j之间的相对位置$\delta(i,j)$定义如下：

$$
\delta(i,j) = \begin{cases} 0 & \text{for} \quad i-j \leqslant -k \\ 2k-1 & \text{for} \quad i-j \geqslant k \\ i-j+k & \text{others} \end{cases}
$$

带有相对位置的解耦自注意力如下公式所示：

$$
Q_c = HW_{q,c}, K_c = HW_{k,c}, V_c = HW_{v,c}, Q_r = PW_{q,r}, K_r = PW_{k,r}
$$

$$
\tilde{A}_{i,j} = \underbrace{\boldsymbol{Q}_i^c \boldsymbol{K}_j^{c\intercal}}_{\text{(a) content-to-content}} + \underbrace{\boldsymbol{Q}_i^c \boldsymbol{K}_{\delta(i,j)}^{r\intercal}}_{\text{(b) content-to-position}} + \underbrace{\boldsymbol{K}_j^c \boldsymbol{Q}_{\delta(j,i)}^{r\intercal}}_{\text{(c) position-to-content}}
$$

$$H_0 = softmax(\frac{\widetilde{A}}{\sqrt{3d}})V_c$$

其中Q_c、K_c、V_c分别是投影内容向量，$P \in R^{2k \times d}$表示跨所有层共享的相对位置嵌入向量（在正向传播期间保持不变）；Q_r和K_r分别是投影相对位置向量。

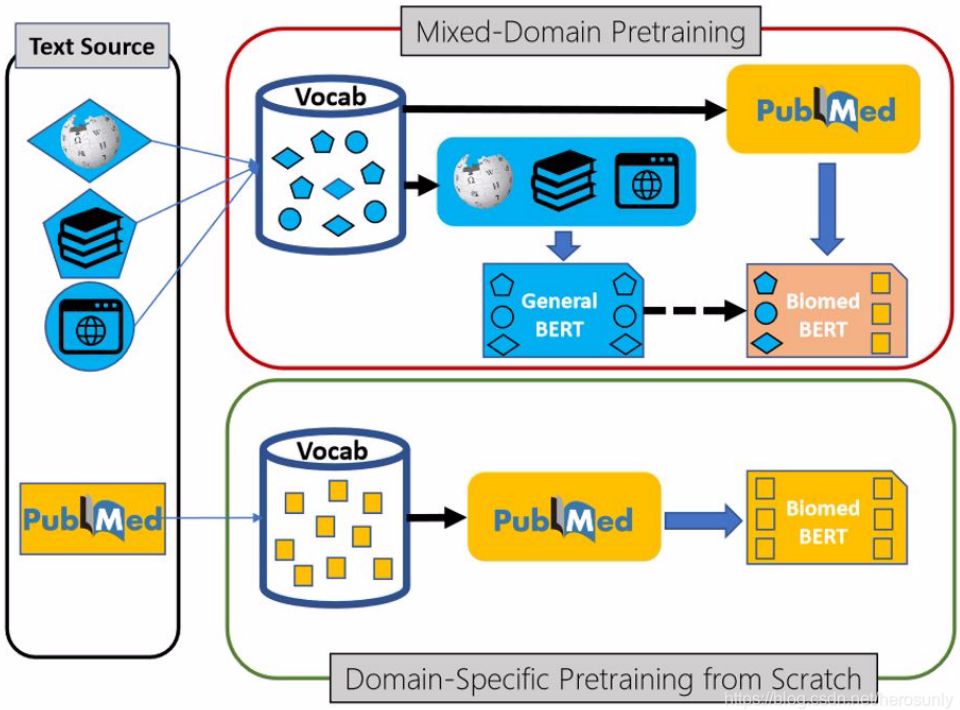
$\widetilde{A}_{i,j}$是注意力矩阵$\widetilde{A}$中的元素，表示tocken i 到j之间的注意力。最后，$\widetilde{A}$还需除以一个缩放系数$\sqrt{3}$，以稳定模型的训练过程。

再来看增强的Masked Decoder。众所周知，原始的BERT存在预训练和微调不一致问题。预训练阶段，隐层最终的输出输入到softmax预测被mask掉的token，而微调阶段则是将隐层最终输出输入到特定任务的decoder。这个decoder根据具体任务不同可能是一个或多个特定的decoder，如果输出是概率，那么还需要加上一个softmax层。为消除这种不一致性，DeBERTa将MLM与其他下游任务同等对待，并将原始BERT中输出层的softmax替换为**「增强后的mask decoder(EMD)」**，EMD包含一个或多个Transformer层和一个softmax输出层。至此，结合了BERT和EMD的DeBERTa成为了一个encoder-decoder模型。

3.2 PubMedBert

PubMedBert是由微软研究人员针对生物医学NLP的领域设计的预训练模型。PubMed是生物医学研究论文的标准存储库，每天增加4,000篇新论文，每年增加一百万篇。

如下图是神经语言模型预训练的两种范例。 图中的上半部分：流行的混合域范例，假设域外文本仍然有用，并且通常使用通用域语言模型初始化特定于域的预训练并继承其词汇。 下半部分：从头开始的针对特定领域的预训练可以导出词汇表，并且仅使用域内文本进行预训练。



研究人员提出了一种新的范例，该范例将从头开始完全只针对特定领域内的域内文本进行预训练BERT。对于生物医学等大批量、高价值的领域，这种策略优于所有先前的语言模型，并在生物医学NLP应用中全面获得了最新技术成果。

在生物医学NLP应用中，PubMedBERT的效果始终优于先前所有的语言模型，而且差距比较大。

	BERT		RoBERTa	BioBERT	SciBERT		ClinicalBERT	BlueBERT	PubMedBERT
	uncased	cased	cased	cased	uncased	cased	cased	cased	uncased
BC5-chem	89.25	89.99	89.43	92.85	92.49	92.51	90.80	91.19	**93.33**
BC5-disease	81.44	79.92	80.65	84.70	84.54	84.70	83.04	83.69	**85.62**
NCBI-disease	85.67	85.87	86.62	**89.13**	88.10	88.25	86.32	88.04	87.82
BC2GM	80.90	81.23	80.90	83.82	83.36	83.36	81.71	81.87	**84.52**
JNLPBA	78.63	78.52	78.81	79.35	79.45	79.56	78.59	78.68	**80.06**
EBM PICO	72.34	71.70	73.02	73.18	73.12	73.06	72.06	72.54	**73.38**
ChemProt	71.86	71.54	72.98	76.14	75.24	75.00	72.04	71.46	**77.24**
DDI	80.04	79.34	79.52	80.88	81.06	81.22	78.20	77.78	**82.36**
GAD	77.72	77.28	77.72	80.94	80.90	79.66	78.40	77.24	**82.34**
BIOSSES	82.68	81.40	81.25	89.52	86.25	87.15	91.23	85.38	**92.30**
HoC	80.20	80.12	79.66	81.54	80.66	81.16	80.74	80.48	**82.32**
PubMedQA	51.62	49.96	52.84	**60.24**	57.38	51.40	49.08	48.44	55.84
BioASQ	70.36	74.44	75.20	84.14	78.86	74.22	68.50	68.71	**87.56**
BLURB score	75.99	75.76	76.33	80.29	78.80	78.08	77.19	76.19	**81.10**

四、实验

4.1 数据集

1. 数据描述

本文提供的数据来自某一医疗执照临床技能考试，目标在于衡量受训人员在遇到标准化病人时识别相关临床信息的能力。在这次考试中，每位经过培训的考生看到一个标准化病人，并对该临床病例进行描述。在与患者交流之后，测试者将患者的相关信息进行记录。每一份病历都由一名训练有素的医生根据所描述的与病例相关的某些关键概念或特征对记录进行评分。

2. 部分术语

（1）临床案例:标准化患者向考生呈现的场景，如症状、投诉、关注事项等。本数据集中有10例临床病例。

（2）病人备注:文字详述了病人在体检和面谈等过程中所涉及的重要信息。

（3）特征:描述了与每个临床案例相关的关键概念。

3. 训练数据

(1) patient_notes数据集

约40000个病历历史部分的集合，该数据集不包括测试集中的患者记录。

关键词	信息
pn_num	病例序号
case_num	科室代码
pn_history	病例的历史记录

(2)features数据集

每个临床病例的特征(或关键概念)的序号及特征的描述信息。

关键词	信息
feature_num	特征序号
case_num	科室代码
feature_text	特征描述

(3) train数据集

病例特征分解的片段及各片段的位置信息。

关键词	信息
id	病例-特征序号
pn_num	病例序号
feature_num	特征序号
case_num	科室代码
annotation	特征片段
location	特征片段所在的位置

4. 测试数据样本

竞赛未给出用于测试模型效果的测试集，当提交模型后，测试集中的患者记录将被添加到patient_notes.csv文件中。这些病人记录与训练集中的病人记录来自相同的临床病例。测试集中大约有2000份病人记录。

test数据集：从训练集中选择的示例实例。

sample_submit数据集：提交文件正确格式的示例。

4.2 数据预处理

1. features数据集

将原“feature_num”为201，“case_num”为2所对应的“feature_text”数据“Last-Pap-smear-I-year-ago”修改为：“Last-Pap-smear-1-year-ago”。

2. train数据集

(1) 横向合并“train”、“features”和“patient_notes”数据集，即在“train”数据集的基础上添加对应的“feature_num”列和“pn_num”列，为新的“train”数据集。

(2) 修改train数据集的“annotation”和“location”中出现的错误，共42条，部分修改记录如下：

原数据：

行	annotation	location
338	'ather heart attack'	'765 783'
621	'or the last 2-3 months'	'78 100'
655	'no old intolerance'	'285 287;297 312'
1262	'mother hyroid problem'	'551 557;566 580'
1396	'stool , with no blood'	'259 271;272 280'

修改后：

行	annotation	location
338	‘father heart attack’	‘764 783’
621	‘for the last 2-3 months’	‘77 100’
655	‘no cold intolerance’	‘285 287;296 312’
1262	‘mother thyroid problem’	‘551 557;565 580’
1396	‘stool , with no blood’	‘259 280’

添加列“annotation_length”，表示“annotation”中涉及的片段个数。

4.3 实验设置

小组使用了两个模型，分别为**“deberta-v3-base”和“BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext”**，预训练模型均来自hugging face。为了使模型的性能不对数据划分方式敏感，小组使用**5折交叉验证**的方式进行模型训练，即每次用80%的数据作为训练集，20%的数据作为验证集。将交叉验证得到的5个模型输入测试集，得到最后一层的输出，我们对输出进行**简单加权平均**得到最后结果，并计算评价指标**micro averaged F1 score**，评价指标的计算方式如下：

$$precision = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} TP_i + \sum_{i=1}^{m} FP_i}, recall = \frac{\sum_{i=1}^{m} TP}{\sum_{i=1}^{m} TP_i + \sum_{i=1}^{m} FN_i}$$

$$F1 = 2\frac{recall \times precision}{recall + precision}$$

其中m是需要分类的tocken个数。和二分类F1值不同的是将所有类的精确率和召回率一起计算，之后再按照F1的公式计算即可。

实验中，小组成员将大部分超参数设置为和BERT原论文相似。设batch_size为4，epoch为5，采用AdamW优化器；对损失函数BCEWithLogitsLoss进行L2正则化，设置weight_decay为0.02；设置max_len为512从而对输入序列的tocken个数进行约束。

根据使用BERT进行fine-tuning的方法，小组成员将数据输入分词器进行分词操作，对于**“deberta-v3-base”**模型，对应的分词器为DebertaV2TokenizerFast，对于**“BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext”**模型，对应的分词器为模型自带分词器。首先，将待训练数据集中的pn_history和feature_text转化为input_ids，token_type_ids和attention_ids，其次，将pn_history结合location得到label，即特殊字符编码为-1，含有特征的段落编码为1，不含特征的段落编码为0。

将input_ids，token_type_ids和attention_ids输入模型，得到最后一个block的输出，在batch_size为4，max_length为512的一次训练中，每批输出的张量维度为（4，512，768）。之后将输出送入一个线性层，将张量映射为（4，512，1），对第2个维度的所有数值计算sigmoid，并设阈值为0.5，将值大于0.5对应的tocken认为是含有特征的段落。

本文所有模型均在Kaggle Notebook中使用GPU进行训练。

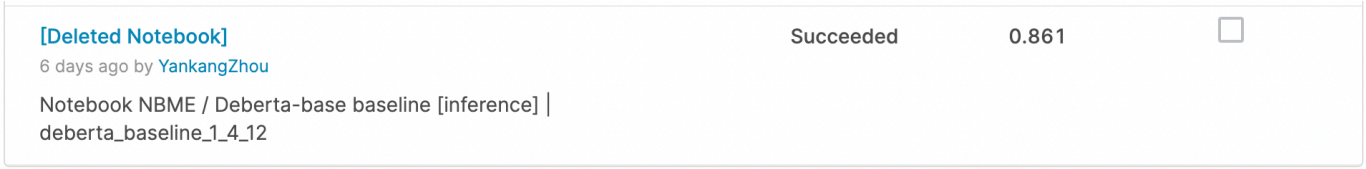
4.4 定量分析

- DeBERTa

本文记录了训练时5折交叉验证中每折模型的训练状态，包括验证指标，训练损失，验证损失如下表。

Fold	Epoch	Val_Score	Loss	Val_Loss
1	5	0.8319	0.0262	0.0137
2	5	0.8569	0.0108	0.0124
3	5	0.8613	0.0094	0.0120
4	5	0.8651	0.0082	0.0119
5	5	0.8648	0.0075	0.0123

可以看到每个fold最好的模型在评估指标上均达到了0.83以上。将五个模型加权的结果送入用户未知的测试集（比赛方用来排名），得到的评估指标**micro averaged F1 score**为0.861，效果达到预期。

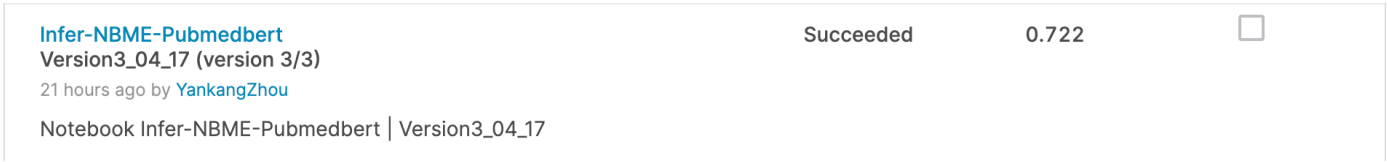


- PubMedBERT

本文记录了训练时5折交叉验证中每折模型的训练状态，包括验证指标，训练损失，验证损失如下表。

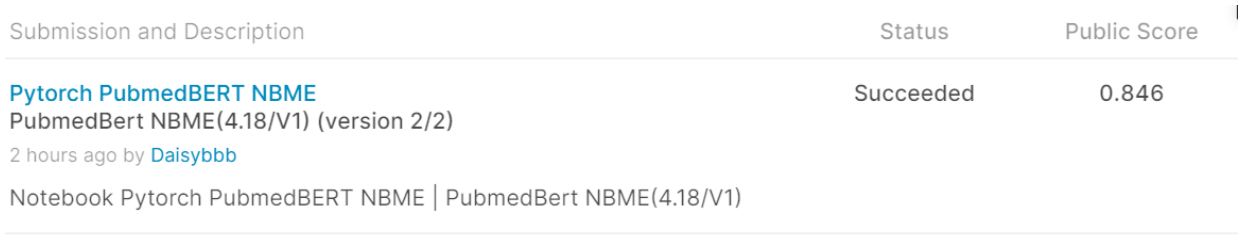
Fold	Epoch	Val_Score	Loss	Val_Loss
1	5	0.7240	0.0076	0.0156
2	5	0.7291	0.0076	0.0158
3	5	0.7212	0.0075	0.0170
4	5	0.7241	0.0075	0.0168
5	5	0.7205	0.0078	0.0167

可以看到每个fold最好的模型在评估指标上均在0.72左右。将五个模型加权的结果送入用户未知的测试集（比赛方用来排名），得到的评估指标**micro averaged F1 score**为0.722，和验证集的效果相似，但最终效果和DeBERTa模型仍有差距。根据验证集损失函数，模型还存在一定的过拟合现象，但由于预训练模型参数过多，无法通过简单调整优化器、dropout等参数解决过拟合。



小组成员又尝试改变训练验证集的划分方式，在上面的实验中，为了方便对比，均使用5折交叉验证，对于每个pn_num，总是划分为同一折，也就是说，某个病例只能出现在训练集或测试集中。这种划分方式是为了让模型真正学习到数据特征而不是根据同一个pn_num的数据推断其他特征段落，但也对模型要求更加严格。

针对PubMedBERT未达预期的现状，使用sklearn的train_test_split简单对训练数据进行8:2的划分，训练5轮后再对模型进行评价，得到0.8153的**micro averaged F1 score**，好于初始数据划分时得到的指标。再将模型送入用户未知的测试集，得到的评估指标为0.846，明显好于该模型上一个版本。



五、结论

针对美国医学执照考试（USMLE）临床技能考试的实际问题，本文将其抽象为QA（或span extraction）问题，它们同属于机器阅读理解的范畴（MRC）。针对该问题的特点，选择使用BERT的变体DeBERTa和PubMedBERT进行fine-tuning。最终分别在模型未见过的测试集上取得了0.861和0.8153的**micro averaged F1 score**值。

对于本文展望，小组成员希望可以解决过拟合问题，同时可以精细化数据预处理和后处理，将指标提高一些。

参考文献

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[3] He P, Liu X, Gao J, et al. Deberta: Decoding-enhanced bert with disentangled attention[J]. arXiv preprint arXiv:2006.03654, 2020.

[4] Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM Transactions on Computing for Healthcare (HEALTH), 2021, 3(1): 1-23.