预测文本的阅读难度

虞月豪 2006100061

李尚谦 2006100078

张伟宏 2006100080

Kaggle 链接: https://www.kaggle.com/c/commonlitreadabilityprize

一、 题目背景介绍

该竞赛是一个确定适合 3-12 年级学生阅读的文本难度级别的比赛。目前，传统的可读性评估方法和商业公式被用于匹配读者和教材，但它们存在一些限制和问题，如文本解码(即每个单词的字符或音节)和句法复杂性（即每个句子的数字或单词）的弱代理，缺乏构建性和理论的有效性，以及成本昂贵。因此，CommonLit, Inc. 与亚特兰大的乔治亚州立大学合作，挑战 Kagglers 们提高文本可读性评估方法。参赛者需要运用机器学习技能和包括不同年龄组读者和来自不同领域的文本的数据集，构建算法来评估阅读材料的复杂性。获胜模型将确保包含文本的连贯性和语义。如果成功，这项挑战将有助于管理员、教师和学生。识字课程开发人员和教师将能够快速准确地评估适合课堂的作品。此外，这些公式将更易于所有人使用。最重要的是，学生将受益于对他们作品的复杂度和可读性的反馈，从而更容易提高他们的阅读能力。

二、 数据集介绍

Files

- **train.csv** - the training set
- **test.csv** - the test set
- **sample_submission.csv** - a sample submission file in the correct format

Columns

- `id` - unique ID for excerpt
- `url_legal` - URL of source - this is blank in the test set.
- `license` - license of source material - this is blank in the test set.
- `excerpt` - text to predict reading ease of
- `target` - reading ease
- `standard_error` - measure of spread of scores among multiple raters for each excerpt. Not included for test data.

数据集分为训练集和预测集，数据项如下：

id - 摘录的唯一 ID

url_legal - 摘录来源的 URL，测试集中为空白。

license - 摘录来源的许可证，测试集中为空白。

excerpt - 需要预测阅读难度的文本摘录

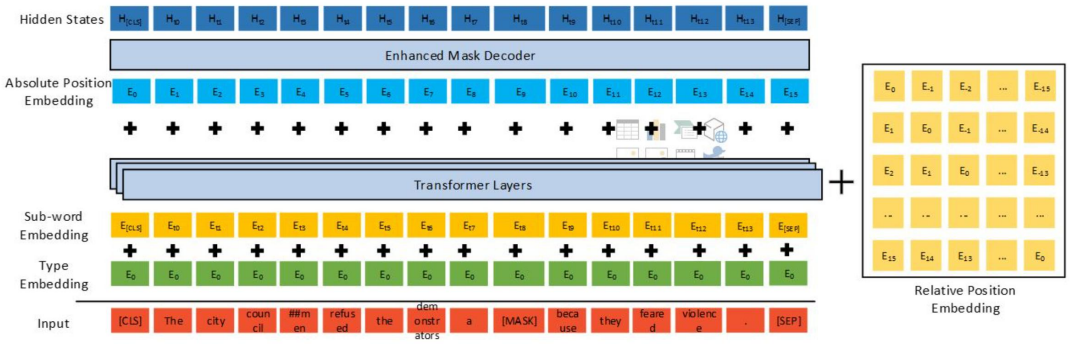
target - 阅读难度

standard_error - 针对每个文本摘录，多个评分人员评分分散程度的度量。测试数据中不包括。

三、 使用的算法介绍

我们计划使用 DeBERTa 模型来预测文本的阅读难度，DeBERTa 原理如下：

DeBERTa（Decoding-enhanced BERT with disentangled attention）使用了注意力解耦机制（下图右边黄色部分）以及增强的掩码解释器(enhanced mask decoder)两种技术对 BERT 和 RoBERTa 模型进行了一个改进，同时还使用了虚拟对抗训练方法来进行模型的改善。



训练步骤如下：

一、 单词注意力权重计算

在 DeBERTa 中其首先会把文本段落拆分成单词，其中每个单词都会使用两个向量来表示其内容与位置，对于序列中位置 i 处的 token，我们使用了两个向量， {Hi} 和{ Pi|j }表示它，它们分别表示其内容和与位置 j 处的 token 的相对位置。 token i 和 j 之间的交叉注意力得分的计算可以分解为四个部分:

$$
\begin{aligned}
A_{i,j} &= \{\boldsymbol{H}_i, \boldsymbol{P}_{i|j}\} \times \{\boldsymbol{H}_j, \boldsymbol{P}_{j|i}\}^\top \\
&= \boldsymbol{H}_i \boldsymbol{H}_j^\top + \boldsymbol{H}_i \boldsymbol{P}_{j|i}^\top + \boldsymbol{P}_{i|j} \boldsymbol{H}_j^\top + \boldsymbol{P}_{i|j} \boldsymbol{P}_{j|i}^\top
\end{aligned}
$$

这样，一个单词对的注意力权重可以使用其内容和位置的解耦矩阵计算为四个注意力（内容到内容，内容到位置，位置到内容和位置到位置）的得分总和。

二、进行 Mask 操作

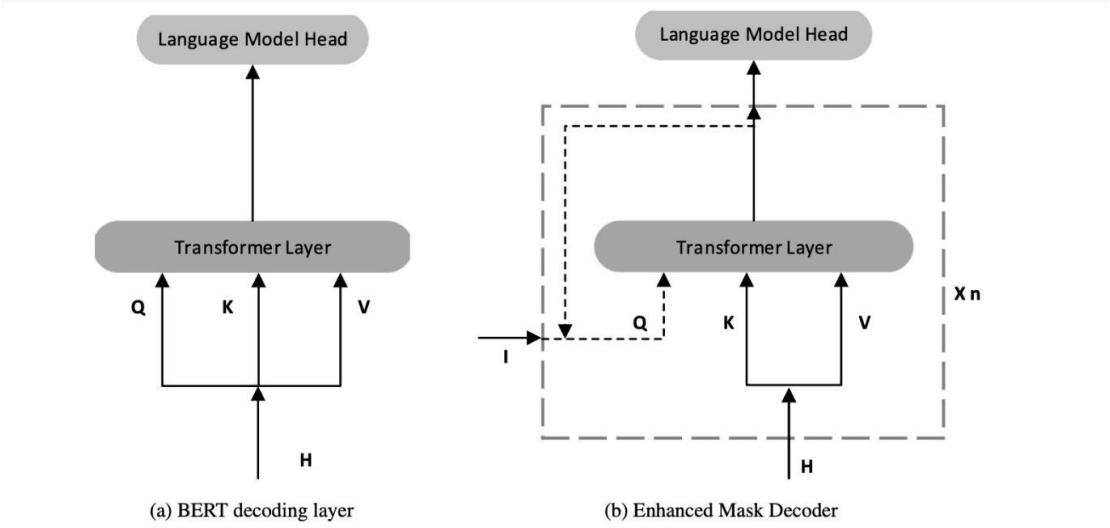
DeBERTa 模型会在一定比例的单词上进行动态 Mask 操作（Masked Self-Attention），使得模型无法直接观察到这些单词，从而训练模型更好地处理未知单词。

三、通过 transformer 层

经过 mask 操作后，这些序列会通过 transformer 层，进行自注意力训练，将输入序列中的每个单词与所有其他单词进行交互，从而获取全局的语义信息。

四、通过增强型掩码解释器

Transformer 解码完成之后，数据会通过 EMD 层，进行绝对位置的合并，并且在 softmax 层之前进行 mask token 预测，如下图所示。这样，DeBERTa 捕获了所有 Transformer 层中的相对位置，同时解码被 mask 的单词时将绝对位置用作补充信息



通过以上步骤之后，基本上就得到了一个 DeBEARTa 模型，随后根据具体的下游任务，输入数据，完成模型的进一步拟合与收敛，这样就能得到一个实现特定自然语言处理任务的模型了。

四、 预计达到的目标

通过使用训练集的数据对 deberta 模型进行微调，来预测测试集中的各文本的阅读难度，结果尽可能接近正确值。

训练集如下：

	id	url_legal	license	excerpt	target	standard_error
1	c12129c31	<null>	<null>	When the young people returned to …	-0.340259125	0.464009046
2	85aa80a4c	<null>	<null>	All through dinner time, Mrs. Fayr…	-0.315372342	0.480804970
3	b69ac6792	<null>	<null>	As Roger had predicted, the snow d…	-0.580117966	0.476676226
4	dd1000b26	<null>	<null>	And outside before the palace a gr…	-1.054013390	0.450007142
5	37c1b32fb	<null>	<null>	Once upon a time there were Three …	0.247197446	0.510844957
6	f9bf357fe	<null>	<null>	Hal and Chester found ample time t…	-0.861808583	0.480936493
7	eaf8e7355	<null>	<null>	Hal Paine and Chester Crawford wer…	-1.759061403	0.476507368
8	0a43a07f1	<null>	<null>	On the twenty-second of February, …	-0.952324620	0.498115881
9	f7eff7419	<null>	<null>	The boys left the capitol and made…	-0.371640688	0.463710362
10	d96e6dbcd	<null>	<null>	One day he had gone beyond any poi…	-1.238432225	0.465899778
11	c57b50918	<null>	<null>	It was believed by the principal m…	-3.081337118	0.553259513

测试集如下：

	C1	C2	C3	C4
1	id	url_legal	license	excerpt
2	c0f722661	<null>	<null>	My hope lay in Jack's promise that he would keep a bright light
3	f0953f0a5	<null>	<null>	Dotty continued to go to Mrs. Gray's every night with the milk.
4	0df072751	<null>	<null>	It was a bright and cheerful scene that greeted the eyes of Cap
5	04caf4e0c	https://en.wiki…	CC BY-SA 3.0	Cell division is the process by which a parent cell divides int
6	0e63f8bea	https://en.wiki…	CC BY-SA 3.0	Debugging is the process of finding and resolving of defects th
7	12537fe78	<null>	<null>	To explain transitivity, let us look first at a totally differe
8	965e592c0	https://www.afr…	CC BY 4.0	Milka and John are playing in the garden. Her little sister is

输出格式如下:

	id	target
1	c0f722661	0.0
2	f0953f0a5	0.0
3	0df072751	0.0
4	04caf4e0c	0.0
5	0e63f8bea	0.0
6	12537fe78	0.0
7	965e592c0	0.0