## 实验任务

使用LCQMC数据集，基于LSTM网络训练一个文本匹配模型。

## 实验设计逻辑

* 1. 关于文本匹配模型：

文本匹配宽泛地说是研究两段文本之间的关系，但在不同的场景下对”匹配“的定义可能非常不同。

文本匹配广泛用于搜索引擎、问答系统、信息流相关文章推荐等场景中，目的是为了在文本库中找出与目标文本类似或相关的文本。

* 1. 程序设计：
     1. 构建数据集
     2. 加载语义模型
     3. 训练模型
     4. 模型预测

## 数据处理

* 1. LCQMC数据集的结构、来源、要求等相关信息

LCQMC: A Large-scale Chinese Question Matching Corpus

LCQMC 是哈尔滨工业大学在自然语言处理国际顶会 COLING2018 构建的问题语义匹配数据集，其目标是判断两个问题的语义是否相同。

LCQMC **口语化描述的语义相似度任务 Semantic Similarity Task**

输入是两个句子，输出是0或1。其中0代表语义不相似，1代表语义相似。

*数据量：训练集(238,766)，验证集(8,802)，测试集(12,500)*

*例子：*

*1.聊天室都有哪些好的 [分隔符] 聊天室哪个好 [分隔符] 1*

*2.飞行员没钱买房怎么办？[分隔符] 父母没钱买房子 [分隔符] 0*

lcqmc数据集 包括训练集,验证集和测试集 问题语义匹配数据集，其目标是判断两个问题的语义是否相同

LCQMC更多的关注在intent matching（**意图匹配**）而不是paraphrase（短语）方面。构建的方式是先针对不同的领域从百度问答中抽取高频的相关问题，然后通过Wasserstein distance进行初步筛选，最后人工进行标注。数据集一共有260068对标注结果，分为三部分，238766训练集、8802验证集和12500测试集。

* 1. 结合LSTM模型确定数据集的处理方法

## 模型设计

### ERNIE语义模型

* 文本本质上是一段语言信息，对于计算机而言，这样的信息是难以直接理解的
* 所以对于NLP任务而言一般需要先将文本转换成数字信息，才能进行信息处理，而这个转换过程一般使用的技术是文本嵌入(Word Embedding)

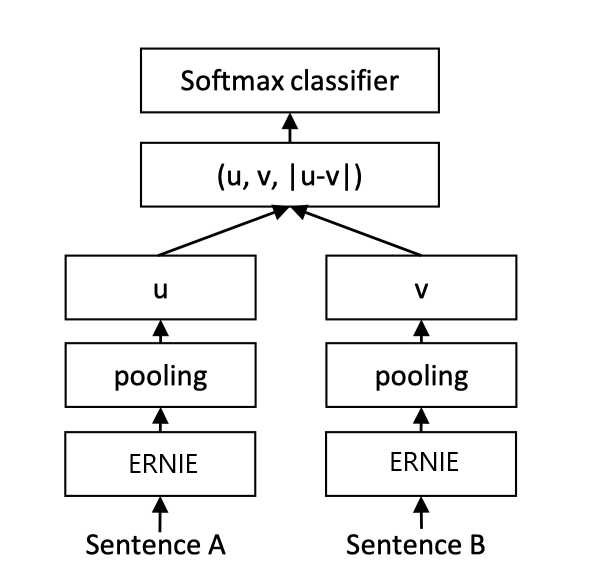


* 文本嵌入的方式有很多，从skipgram、cbow这类的word2vec模型，发展到后来的ELMo模型，再到现在大量使用的基于预训练的ERNIE、BERT模型等等
* 相比之前的模型，现在的ERNIE、BERT等等这些模型由于引入大量的真实预料进行了有效的预训练，所以能够更好的完成文本嵌入的工作
* 综上所述，ERNIE一般是作为完整模型的一个组件，作用一般为文本嵌入(Word Embedding)，即将输入文本转换成向量表示输出

### 文本相似度判断

使用Pointwise文本匹配模型完成文本相似度判断：

1. 将输入的两个文本分别通过同一个ERNIE模型得到其对应的嵌入向量
2. 对两个嵌入向量和他们之间的差值进行拼接
3. 对拼接向量进行线性转换后进行二分类，即做一个是否相似的二分类
4. 结构图如下：



### PaddleHub数据集构建



## 训练配置

### 超参数

##### 学习率

    learning\_rate=5e-5,

##### 语义模型

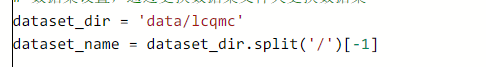
ernie

##### 网络

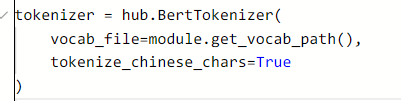
    network='lstm',

### 实例化

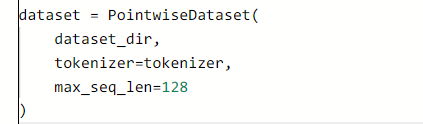
#### 数据集设置



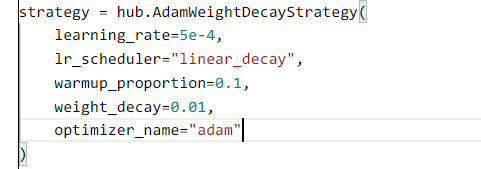
#### 加载tokenizer



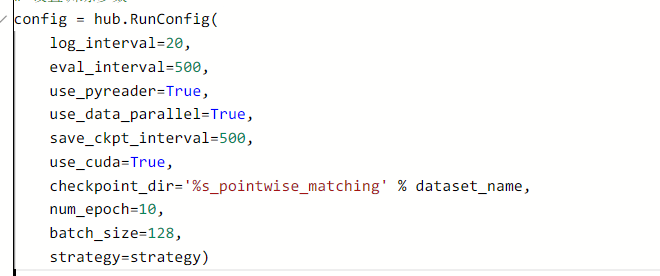
#### 定义数据集



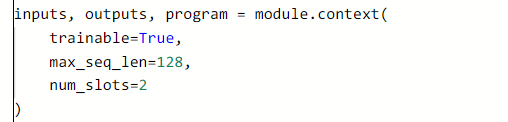
#### 设置优化策略



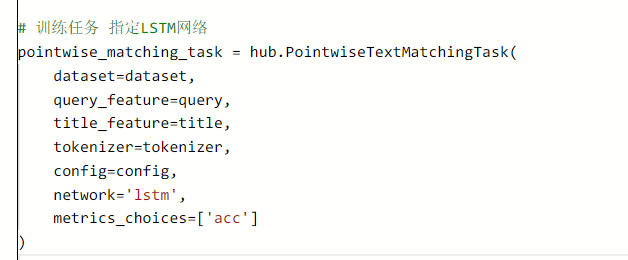
#### 设置训练参数



#### 定义模型上下文



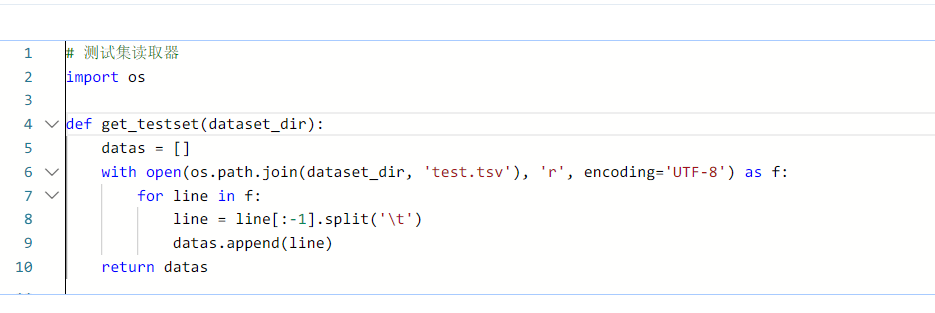
#### 设置训练任务

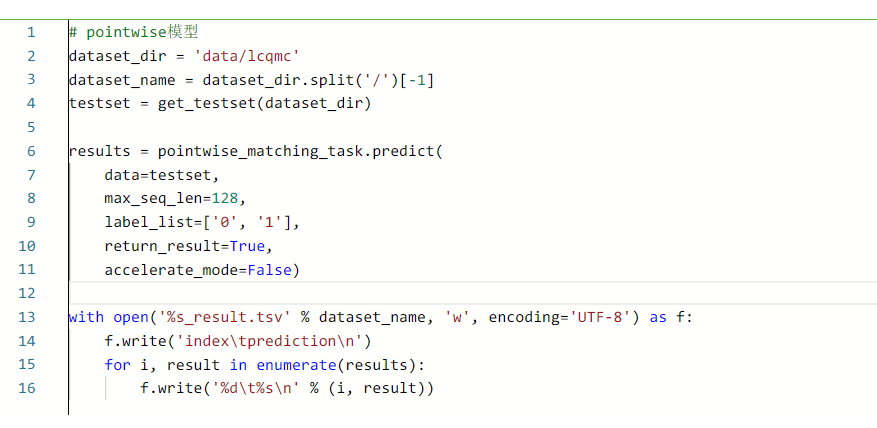


#### 开始训练



#### 模型预测





### 硬件资源



### 优化策略

# 设置优化策略

strategy = hub.AdamWeightDecayStrategy(

    learning\_rate=5e-4,

    lr\_scheduler="linear\_decay",

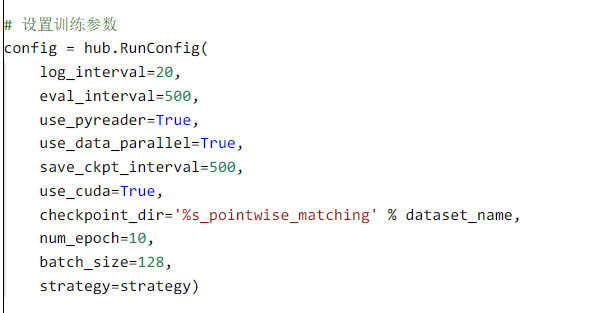
    warmup\_proportion=0.1,

    weight\_decay=0.01,

    optimizer\_name="adam"

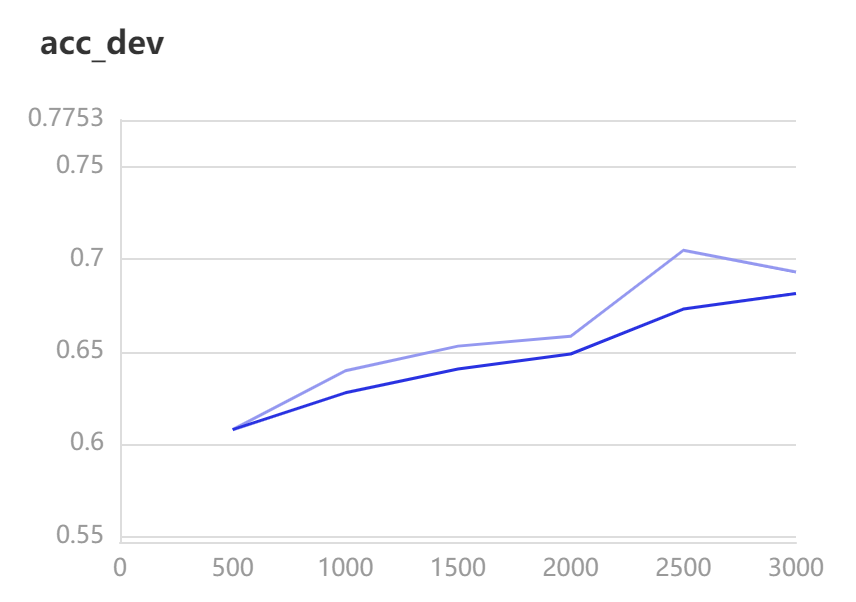
)

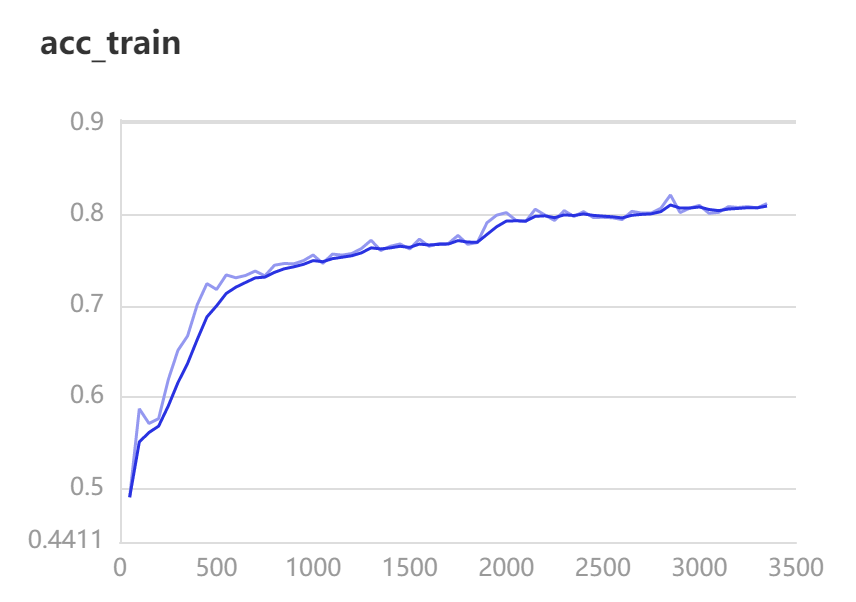
### 训练参数

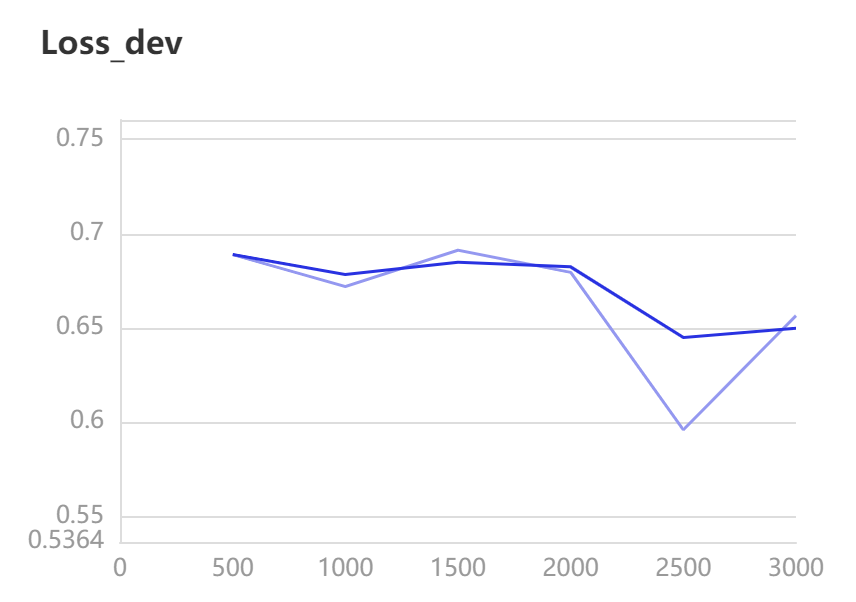


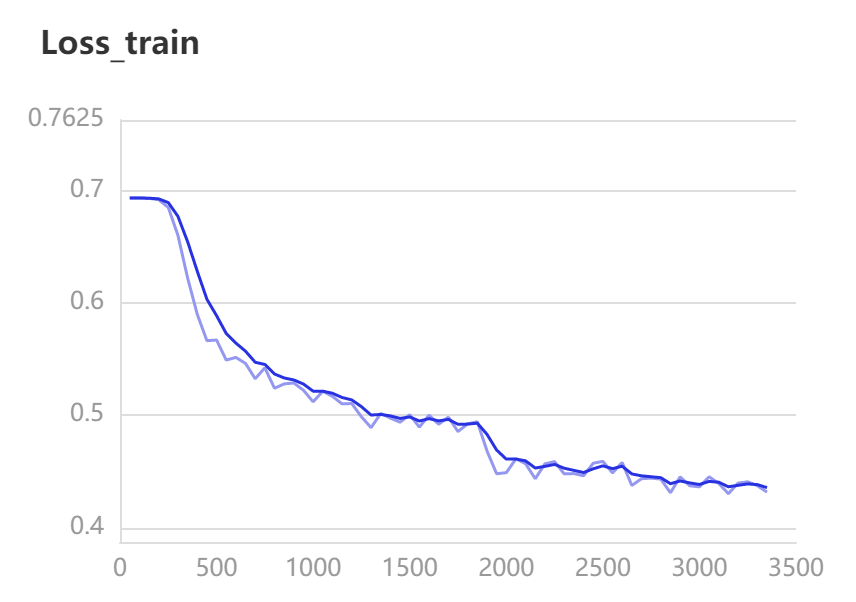
## 模型训练与评估

### 训练过程

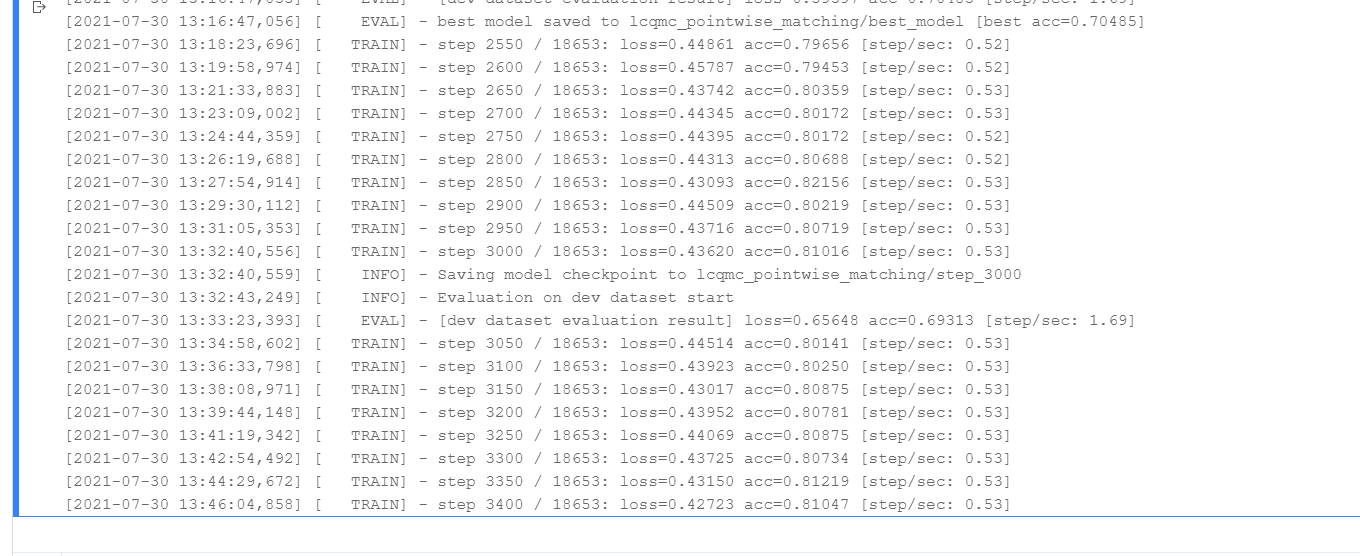




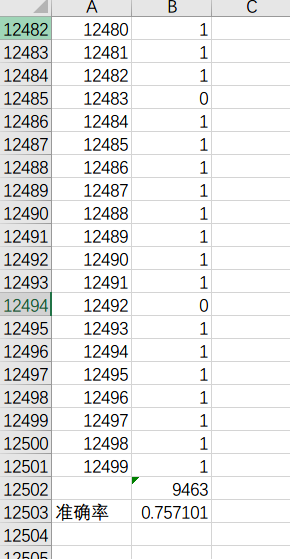




由于时间关系，只训练了1/5的训练数据



### 模型评估



## 模型推理

通过传入PointwiseDataset的dataset\_dir，可以指定相同数据格式的数据

