1. 赛题理解

信息流场景下，短视频消费引来爆发式增长。本赛题主要利用多模态机器学习技术来探索视频 Embedding 的构建。可以通过离线训练的方式得到给定视频的 Embedding 向量，通过计算视频 Embedding 之间的余弦相似度，采用 Spearman's rank correlation 与人工标注相似度计算相关性，并最终排名。

1.1. 数据集介绍

- Pointwise 数据集：主要是百万级单点视频数据，包括视频 ID，已经抽取好的 frame 特征，视频标题文本，视频 ASR 文本，人工标注的视频 TAG，视频的分类；
- Pairwise 数据集：主要包括万级别的单点视频数据，包括视频 ID，已经抽取好的 frame 特征，视频标题文本，视频 ASR 文本，人工标注的视频 TAG，视频的分类；组委会提供了在 Pairwise 数据集上的 Pairwise label 文本，包括 Query 视频 ID，Candidate 视频 ID，以及人工标注的相似度；
- Test 数据集：主要包括万级别的单点视频数据，包括视频 ID，已经抽取好的 frame 特征，视频标题文本，视频 ASR 文本；

1.2. 数据分析

- 人工标注数据 TAG 主要有几万个，baseline 算法当前只提供了 1 万个 TAG 列表可以作为预训练的标注信息。人工标注的类别信息，前三位是一级分类。
- 人工对视频相似度打分主要分为 21 个类别，其中大部分分数在 0.2 分左右。经过观察分析人工标注 TAG 和分类信息与评分之间没有发现明显的强关联规则。
- Spearman's rank correlation 存在 rank 误差，采用 Embedding 很难保证不同的两个视频 Embedding 计算出相同的分数。

1.3. 提交结果

- 在 Test 数据集上，提交结果 json 文件，主要包括视频 ID，已经对应的 Embedding 向量，最大长度是 256 维，命名为 result.json，采用 zip 压缩包方式提交。

1.4. 运行环境

- 机器配置：显卡 4*V100；内存 256G；硬盘 2T
- Python：3.8.11
- Python 依赖：

---

tensorflow-gpu==2.3.1

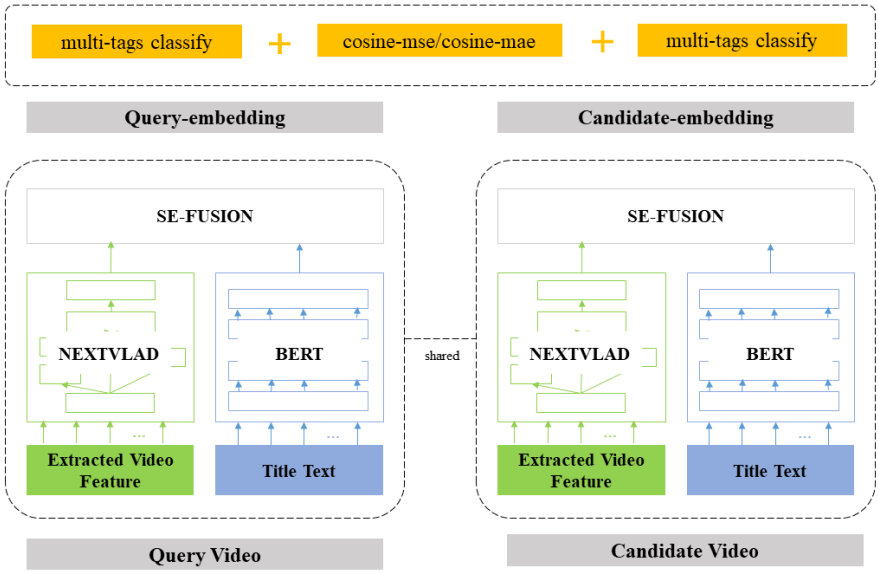
transformers==3.2.0
scikit-learn
tqdm
jieba
pandas
numpy
scipy

2. 模型介绍

本次比赛，我们设计并实现了三种类型的模型，其中第三种模型效果训练单模效果最好，五折结果在测试集 B 的表现为 0.8159，模型一和模型二的五折结果在测试集 B 的表现为 0.808。

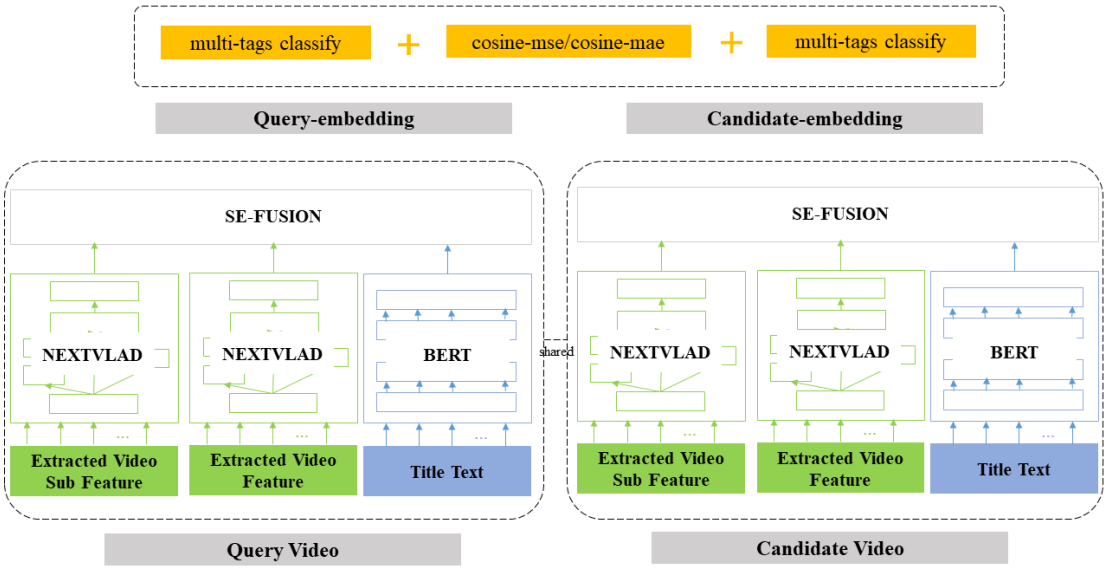
2.1. 模型一

- 基于 baseline 模型修改，输入支持两个视频，经过一个完全共享权重双塔模型后可以输出两个视频对应的 256 维的向量
- 双塔模型内部支持两个模态，其中已抽取的视频特征经过 NEXTVLAD 结构可以得到视频特征的表征 Embedding，标题文本经过 BERT 结构可以得到 title 的表征 Embedding
- 双塔模型内部直接将得到的视频特征的表征 Embedding 和 title 的表征 Embedding 进行 SE 融合拼接，得到最终的多模态视频 Embedding
- 网络采用多 Loss 设计实现，包括 Query 视频在标注 tag 上的多标记分类 loss、Candidate 视频在标注 tag 上的多标记分类 loss，已经 Query 视频和 Candidate 视频在计算 cosine 距离之后和人工标注分数的 mse 和 mae 损失。
- 将所有 Loss 相加之后计算整体损失，针对 BERT 和其他网络参数分别采用两种优化策略来进行优化。



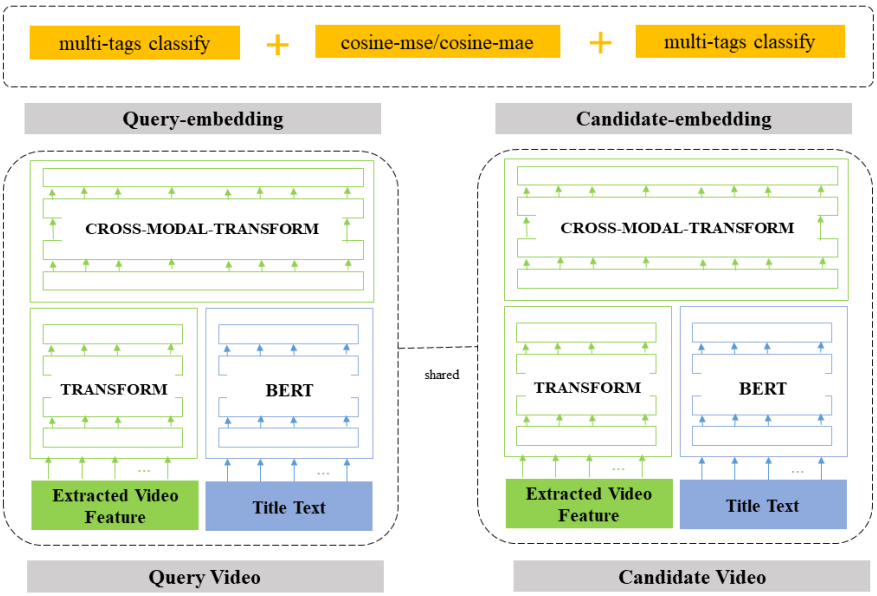
2.2. 模型二

- 基于模型一修改，输入支持两个视频，经过一个完全共享权重双塔模型后可以输出两个视频对应的 256 维的向量
- 双塔模型内部支持三个模态，新增加相邻两个帧特征的帧差表达，我们认为相邻帧之间的表达应该比较接近，如果出现大的转场镜头等应该是一个比较强的特征
- 其它都和模型一保持一致



2.3. 模型三

- 基于模型一修改，输入支持两个视频，经过一个完全共享权重双塔模型后可以输出两个视频对应的 256 维的向量
- 双塔模型内部支持两个模态，其中已抽取的视频特征经过 TRANSFORM 结构可以得到视频特征的表征 Embedding，标题文本经过 BERT 结构可以得到 title 的表征 Embedding
- 双塔模型内部直接将得到的视频特征的表征 Embedding 和 title 的表征 Embedding 进行 CROSS-MODAL-TRANSFORM 融合，得到最终的多模态视频 Embedding
- 其它都和模型一保持一致



3. 关键点介绍

- 首先在 pointwise 数据集上进行预训练，在双塔模型中，query 和 candidate 视频都采用相同的视频进行输入，最终模型只是在 tag_list 上实现多标记训练
- 接着在 pairwise 数据集上进行 fine-tuning，在双塔模型中，query 和 candidate 视频都可以输入到网络中，在训练 query 和 candidate 的 tag_list 多标记分类的同时，计算 query 和 candidate 余弦相似度和人工相似度的 mse/mae
- 因为 tag_list 分布不均匀，所以我们在计算 mse/mae 的时候对人工标注分数进行了变换，变换公式如下(α是人工设定超参数)：

$$\text{changed}(score) = \text{sigmoid}(\alpha * score)$$

4. 模型融合

4.1. 策略一

初赛使用融合策略：直接将所有输出的结果进行正则化之后相加平均：

$$final = \frac{1}{N}\sum_{i=0}^{N} \mathrm{L2}(results_i)$$

4.2. 策略二

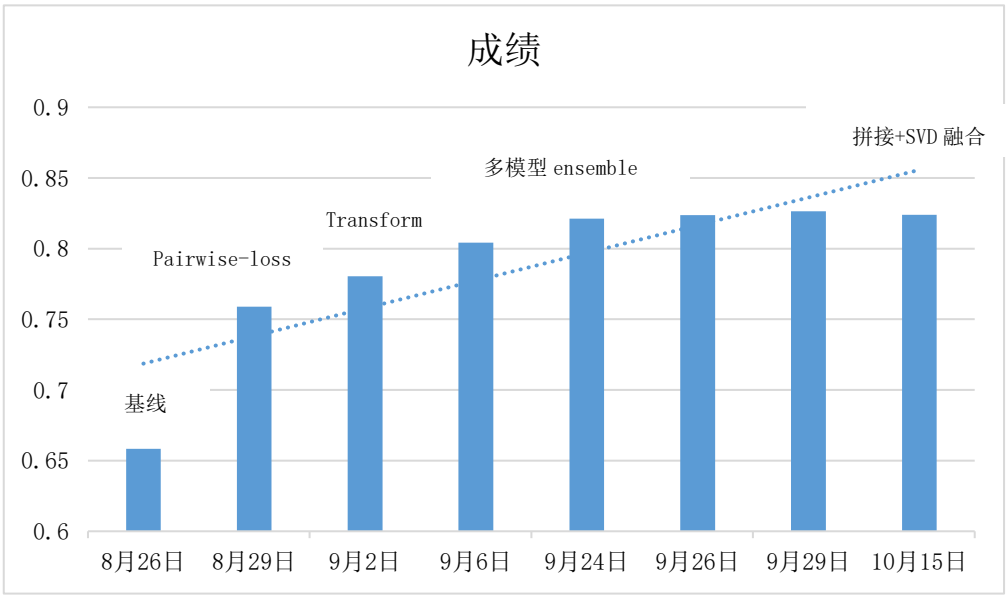
复赛使用融合策略：直接将所有输出的结果进行正则化之后拼接，采用 arpack svd 降维的方式降维到 256 维：

$$final = \mathrm{SVD}_{arpack}(\underbrace{\mathrm{concat}(\mathrm{L2}(results_i))}_{for\ all\ i\ results})$$

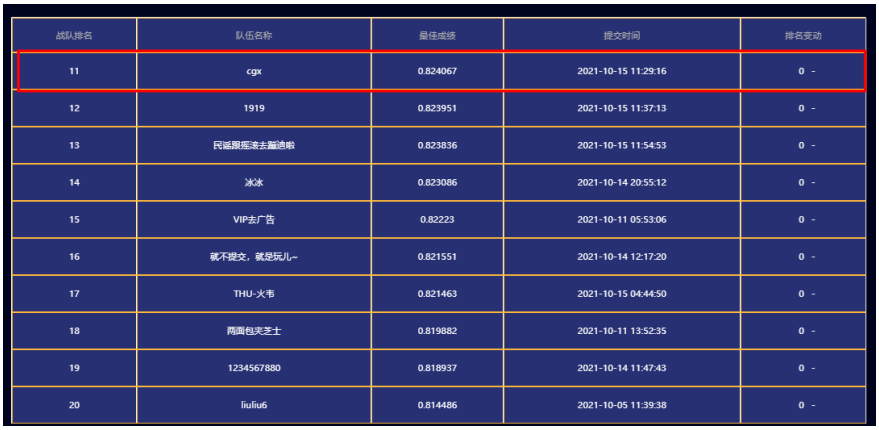
5. 总结思考

在信息流推荐领域，视频发挥着至关重要的作用，这次比赛通过切身实践让我们了解到了如何做好视频内容的加工、理解，更好的服务于上层的业务，从而带来产品竞争力的提升与团队凝聚力的提升。

本次比赛的融合提分点：



本次比赛排名：



战队排名	队伍名称	最佳成绩	提交时间	排名变动
11	cgx	0.824067	2021-10-15 11:29:16	0 -
12	1919	0.823951	2021-10-15 11:37:13	0 -
13	民谣跟摇滚去酿造啦	0.823836	2021-10-15 11:54:53	0 -
14	冰冰	0.823086	2021-10-14 20:55:12	0 -
15	VIP去广告	0.82223	2021-10-11 05:53:06	0 -
16	就不提交，就是玩儿~	0.821551	2021-10-14 12:17:20	0 -
17	THU-火韦	0.821463	2021-10-15 04:44:50	0 -
18	两面包夹芝士	0.819882	2021-10-11 13:52:35	0 -
19	1234567880	0.818937	2021-10-14 11:47:43	0 -
20	liuliu6	0.814486	2021-10-05 11:39:38	0 -

6. 致谢

感谢组委会提供这样一次宝贵的学习与交流机会，感谢组织方的耐心答疑
感谢公司领导的支持，感谢队友的无私付出与努力
感谢大家的帮助与支持

引用

[1] qq 浏览器 2021 年算法大赛：https://algo.browser.qq.com/
[2] transformer：https://github.com/huggingface/transformers
[3] svd：https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html