中国高校计算机大赛

2022 中国高校计算机大赛 —— 微信大数据挑战赛

通　知

2016 年，教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会、全国高等学校计算机教育研究会联合创办了“中国高校计算机大赛”（China Collegiate Computing Contest，简称 C4），目前“中国高校计算机大赛”继续由全国高等学校计算机教育研究会主办。大数据挑战赛是其中的一项重要赛事，在 2018-2021 年均入选全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2022 中国高校计算机大赛——微信大数据挑战赛（以下简称“大赛”）是由清华大学和腾讯微信事业群联合举办，腾讯云提供竞赛平台和资源支持，以企业真实场景和实际脱敏数据为基础，面向全球开放的高端算法竞赛。大赛旨在通过竞技的方式，提升人们对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用。

本次大赛面向全球开放，不限年龄国籍，高等院校在校学生（包括高职高专、本科生、研究生）以及科研机构和企业从业人员均可报名参赛。参赛队伍根据赛题要求设计相应的算法进行数据分析和处理，比赛结果按照指定的评价指标使用在线评测数据进行评测和排名，得分最优者获胜。

请各学校积极配合，按照通知和大赛章程做好宣传和组织工作，为在校生和毕业生参与竞赛提供必要的条件和支持。

竞赛详情见附件（2022 大数据挑战赛竞赛规程）。

全国高等学校计算机教育研究会

2022 年 4 月

2022 中国高校计算机大赛——微信大数据挑战赛
竞赛规程

2016 年，教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会、全国高等学校计算机教育研究会联合创办了“中国高校计算机大赛”（China Collegiate Computing Contest，简称 C4），目前“中国高校计算机大赛”继续由全国高等学校计算机教育研究会主办。大数据挑战赛是其中的一项重要赛事，在 2018-2021 年期间均入选全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2022 中国高校计算机大赛——微信大数据挑战赛（以下简称“大赛”）由清华大学和腾讯微信事业群联合举办，由腾讯云提供大赛资源支持。本次大赛是以企业真实场景和实际脱敏数据为基础、面向全球开放的高端算法竞赛。大赛旨在通过竞技的方式，提升人们对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用。

一、参赛对象

本次大赛面向全球开放，不限年龄国籍，高等院校在校学生（包括高职高专、本科、研究生）以及科研机构和企业从业人员均可参赛。具体要求如下：

- 可以自由组队参赛，具体组队要求见报名&组队的相关说明；
- 参赛选手应保证报名信息真实准确有效，如队伍中的选手信息不符合要求，组委会有权取消整个队伍的参赛资格及奖励。

为了保证大赛的公平性，大赛主办和技术支持单位如有机会接触赛题和相关数据的人员不允许参赛。

腾讯公司内部非接触赛题和相关数据的人员可报名参赛，具体规定如下：

- 实习生不受限制，以所在学校报名参赛；
- 参赛队伍中如有一位腾讯公司正式员工，即视为腾讯公司参赛队伍；
- 是否正式员工以大赛启动报名时间的状态为准；
- 腾讯公司参赛队伍的成绩可参与排名，入围复赛队伍数量不超过 10 支，不允许入围决赛以及获得大赛奖金，可获得比赛名次证书。

二、赛制说明

本次大赛分为报名&组队、初赛、复赛和决赛等四个阶段，其中初赛阶段由参赛队伍下载数据在本地进行算法设计和调试，并通过大赛报名官网提交结果文件；复赛阶段要求参赛队伍在大赛官网平台上进行数据处理、算法调试和生成结果，数据不可下载，可使用平台提供的计算资源和工具包；决赛要求参赛者进行现场演示和答辩。

1. 报名&组队（4 月 26 日 – 6 月 21 日）

参赛选手须在大赛官网或小程序“微信大数据挑战赛”上报名并且组队参赛（即使单人参赛也要组建单人队伍），大赛不收取任何报名费用。

大赛报名小程序
“微信大数据挑战赛”

大赛报名系统开放时间为北京时间 2022 年 4 月 26 日 10:00，截止时间为北京时间 2022 年 6 月 21 日中午 12:00。

- 报名方式：登录比赛官网，完成个人信息注册，即可报名参赛；
- 每个选手可单人成队或 2-3 人组队参赛，且每人只能参加一支队伍。

大赛官方渠道主要包括：

- 大赛官网：https://algo.weixin.qq.com/
- 大赛小程序：微信大数据挑战赛
- 大赛公众号：微信大数据挑战赛
- 大赛邮箱：data@tsinghua.edu.cn
- 大赛 QQ 群：762146461 / 901317172

报名截止之后，不再允许添加或更改任何队伍成员。如有中途退出情况，只允许在参赛队伍内部更换队长或删除队员。参赛队伍须应在决赛开始前由队长和涉及变更的队员向大赛组委会共同提交变更申请，经由大赛组委会审核确认后变更生效。

2. 初赛（5 月 20 日 – 6 月 22 日）

参赛队伍可从大赛官方网站下载数据，在本地进行算法调试，并在线提交结果。

5 月 20 日 10:00 开始，选手可以从竞赛平台下载初赛训练数据集，用于参赛队伍训练模型以及制定预估策略；同时，平台提供测试数据集，用于参赛队伍在比赛中的模型评估和排名。

初赛采用 AB 榜形式：

- 初赛 A 阶段：5 月 20 日 10:00 – 6 月 21 日 20:00，每个参赛队伍每天可以有 3 次提交结果机会，系统实时评测并返回成绩。排行榜每小时更新，将选择参赛队伍在本阶段的历史最优成绩，按照评测指标从高到低排序。
- 初赛 B 阶段：6 月 22 日 10:00-20:00。系统将在 6 月 21 日 21:00 更换测试数据，参赛队伍需再次下载数据文件。本阶段提供 2 次提交结果的机会，系统进行实时评测并返回成绩。排行榜每小时进行更新，并选择参赛队伍在本阶段的

历史最优成绩进行排名展示。

初赛提交的截止时间是 6 月 22 日 20:00，初赛以 B 榜成绩作为初赛成绩依照，要求 TOP110 团队提交代码审核，代码提交截止时间是 6 月 26 日中午 12:00。

组委会将审核并取消存在人工标注、相互抄袭等行为队伍的比赛资格，晋级空缺名额后补。对于初赛成绩符合要求且通过实名认证的参赛队伍，排名前 75 名的参赛队伍将进入复赛，所有通过审核的队伍将获得初赛名次证书。

3. 复赛（7 月 1 日 – 8 月 5 日）

复赛阶段测试数据不可见且不可下载，采用 docker 镜像的方式进行提交，由选手提交打包好的代码镜像来运行得出预测结果，并对时间复杂度有限制。具体要求和说明见“容器镜像”文档（复赛开始前在大赛网站公布）。

预热阶段（7 月 1 日 12:00 – 7 月 4 日 12:00）：系统每天提供 5 次实时评测供选手进行提交测试，排行榜不展示排名情况。

正式阶段（7 月 5 日 12:00 – 8 月 5 日 22:00）：复赛采用 AB 榜形式。

- 复赛 A 阶段：7 月 5 日 12:00 – 8 月 4 日 12:00，每个参赛队伍每天可以有 2 次提交结果机会，系统实时评测并返回成绩。排行榜每小时更新，将选择参赛队伍在本阶段的历史最优成绩，按照评测指标从高到低排序。
- 复赛 B 阶段：8 月 4 日 18:00 – 8 月 5 日 22:00，系统将在 8 月 4 日 18:00 提供 B 榜测试数据集。本阶段提供 2 次提交结果的机会，系统实时评测并返回成绩。排行榜每小时更新，将选择参赛队伍在本阶段的历史最优成绩，按照评测指标从高到低排序。

复赛截止时间是 8 月 5 日 22:00，复赛 TOP30 团队需提交代码审核。组委会将审核并剔除只靠人工标注而没有算法贡献的队伍，晋级空缺名额后补，最终通过复赛成绩审核的前 6 名队伍将晋级决赛。

4. 决赛（8 月下旬）

决赛将以现场答辩会的形式进行，具体要求和安排另行通知。受邀参加决赛的选手在决赛期间的食宿由大赛组委会安排，往返交通费及其他费用自理。

晋级决赛团队需提前准备答辩材料，包括路演 PPT、参赛总结、算法核心代码。在决赛答辩会上，每支队伍面对评委有 20 分钟的路演时间和 10 分钟的答辩时间。评委将根据选手的技术思路、理论深度、算法性能和现场表现进行综合评分。

决赛分数将根据参赛队伍的算法成绩和答辩成绩加权得出，评分权重为复赛阶段 70%，决赛答辩 30%。

三、奖项设置

1. 初赛奖项

初赛 TOP110 且通过代码审核的团队将颁发初赛名次证书，此项奖励以大赛官网初赛最终排行榜为准。

2. 复赛与决赛奖项

大赛奖金池总额为 56 万元人民币，所有奖金均为税前金额，此项奖励以大赛官网复赛最终排行榜和决赛结果为准。

奖励对象	数量	奖励办法
决赛第 1 名队伍	1	奖金 30 万元，决赛名次证书
决赛第 2 名队伍	1	奖金 10 万元，决赛名次证书
决赛第 3 名队伍	1	奖金 6 万元，决赛名次证书
决赛第 4-6 名队伍	3	奖金 2 万元，决赛名次证书
复赛第 7-10 名队伍	4	奖金 1 万元，复赛名次证书
复赛第 11-30 名队伍	20	复赛名次证书

3. 在校学生队伍奖项

在校学生队伍要求所有参赛队员必须全部为在校学生，如果队伍中有一名在职人员，则整个队伍视为在职人员队伍。其中中国大陆在校学生提供学信网的教育部学籍在线验证报告编号进行身份验证，其余学生提供相关在读证明进行身份验证，在校学籍以大赛报名时间 2022 年 4 月 26 日为准。

此奖项仅颁发给进入复赛的在校学生队伍，要求队伍提交复赛代码并获得有效的复赛成绩，根据所有在校学生队伍复赛成绩的单独排名结果进行颁发。

奖项名称	数量	对象
全国一等奖	10	单独排名第 1-10 名
全国二等奖	20	单独排名第 11-30 名
全国三等奖	N	单独排名第 31 名及之后的队伍

另外，如果入围决赛的队伍中没有在校学生队伍，大赛将增补一个决赛名额，提供给单独排名第一名的在校学生队伍入围决赛。

4. 周周星

在初赛阶段，设立周周星奖励。从初赛第三周开始，以每周一中午 12 点的排行榜为准，取前两名参赛队伍发放周周星纪念礼物；另外视情况会额外增加一个名额，以保证当期周周星队伍中至少有一个在校生队伍或一个在职人员队伍。对于前面已经获得周周星的队伍，不重复发放，名额按名次顺延。

四、违规处理

参赛者应本着诚实、公平的态度参加比赛，如在以下情况出现违规，大赛组织委员会（简称“组委会”）有权取消参赛者所在队伍的参赛资格，情节严重者将通报参赛者所在高校并追究其违法责任。

1. 账号使用：参赛者有义务保证账号信息的真实性和有效性，且账号仅限于参赛者本人使用；参赛者禁止使用多账号参赛，同一参赛者不可使用多个账号进行提交、刷分操作；如根据判断认为参赛账号存在异常或违背正常使用条例，组委会可以单方面暂停或终止该账号登录大赛平台。
2. 比赛成果：
    - 严禁参赛队伍之间相互抄袭。如不同参赛队伍提交结果高度相似，经判定存在抄袭行为的，组委会将取消相关参赛队伍的参赛资格，相关参赛成绩无效。
    - 参赛者应保证其在比赛过程中所产出的所有成果未侵犯任何第三方的知识产权、商业秘密及其他合法权益。如第三方因为参赛者侵权行为提出索赔、诉讼等，参赛者应承担由此产生的全部责任及损失。
    - 如大赛主办方及其关联公司有意取得参赛者在本次大赛中独立开发的依约定享有完整知识产权的研究成果，参赛者同意大赛主办方及其关联公司在同等条件下享有优先受让权，相关转让事宜由双方另行协商确定。
3. 数据使用：对于大赛提供的数据（数据集），参赛者须仅在比赛场景下使用，并应妥善保存已下载的数据（数据集），避免泄露；在完成比赛使用后应及时销毁已下载数据（数据集）；如使用比赛之外的任何数据应获得组委会许可。对于不提供下载的比赛数据，参赛者不得以任何形式擅自复制、下载或获取。参赛者如发现任何出现数据未授权访问的可能，应立即通知组委会并积极提供相关信息。如参赛者泄露已下载的数据（数据集），或未及时销毁已下载的数据（数据集）导致已下载的数据（数据集）泄露，参赛者应承担由此产生的全部责任及损失。
4. 代码分享：在大赛举办期间，未经组委会同意，参赛者禁止公开分享与赛事相关的数据、模型和代码；大赛结束之后，参赛者可以在拥有模型和代码的知识产权的情况下自行选择公开分享，但需要确保此类公开共享不会侵犯任何第三方的知识产权、商业秘密及其他合法权益。

5. 参赛者若在参赛过程中发现相关规则漏洞或技术漏洞，有义务及时告知资委会相关漏洞的信息，组委会将对提供相关信息的参赛者表示相关感谢；若参赛者利用相关漏洞进行参赛，经判断查证后，成绩将会被判断为无效成绩。

五、申诉与仲裁

1. 参赛团队或选手对不符合大赛规定的设备、工具和软件，有失公正的评判和奖励以及工作人员的违规行为等，均可向大赛组委会提出申诉。组委会负责受理比赛中提出的申诉并进行调解仲裁，以保证大赛的顺利进行和大赛结果的公平公正。组织委员会作出的仲裁结果为终局决定。
2. 申诉报告应明确申诉内容，指定一名成员作为联系人，通过大赛邮箱以邮件发送，否则申诉将不予以受理。
3. 组织委员会将在收到申诉之日起 5 个工作日之内受理，并认真核查和处理。

六、其他说明

1. 为了确保整个大赛顺利、公正地进行，以及保证参赛选手的合法权益，参赛选手报名时应阅读和确认大赛官网上的《参赛协议》，自觉遵守协议规定。
2. 在大赛举办过程中，竞赛规程可能会有少量的变更和调整，大赛组委会将本着公平、公正、公开的原则在大赛官网公告，所有内容均以大赛官网为准。

“中国高校计算机大赛——微信大数据挑战赛”组织委员会

2022 年 4 月

附件：赛题描述 —— 多模态短视频分类

多模态短视频分类是视频理解领域的基础技术之一，在安全审核、推荐运营、内容搜索等领域有着十分非常广泛的应用。一条短视频中通常包含有三种模态信息，即文本、音频、视频，它们在不同语义层面的分类体系中发挥着相互促进和补充的重要作用。微信产品的内容生态繁荣，创作者覆盖范围大，导致短视频数据中普遍存在着模态缺失、相关性弱、分类标签分布不均衡等问题，是实际应用中需要着重解决的技术难点。本赛题要求参赛队伍基于微信视频号短视频数据以及对应的分类标签标注，采用合理的机器学习技术对指定的测试短视频进行分类预测。

一、 竞赛数据

比赛分为初赛和复赛两个阶段：初赛阶段提供百万量级的无标注数据和十万量级的有标注数据用于训练；复赛阶段训练数据和初赛相同，**主要区别是初赛阶段只提供视频抽帧特征，而复赛阶段提供视频抽帧原始图像。**初赛阶段所有训练数据对参赛队伍开放下载；复赛阶段的训练数据为闭源数据，参赛队伍在**腾讯云 TI-ONE 平台**完成训练。

1. 数据格式

字段名	类型	举例	说明	备注
id	string	13655102198344648800	视频唯一 ID	
category_id	string	2117	人工标注的视频分类 ID	category_id 固定为 4 位字符：前两位为一级分类 ID，后两位对应一级分类下的二级分类 ID。
title	string	苏炳添刷新亚洲记录小组第一轻松晋级百米决赛#奥运@微信时刻	视频标题	可能存在空值。
frames_feature	float list	[[0.89, 1.86, -4.67, -4.38, ...], [0.13, 1.11, -2.12, -3.24, ...],]	视频帧的特征	使用预训练模型提取的视频帧特征。每秒抽取一帧进行提取。每个视频最多提供前 32 帧的特征，超出的部分不会被使用。
frames	string	13655102198344648800.zip	视频帧打包的路径	视频帧的原始图像。每秒抽取一帧。每个视频最多提供前 32 帧图像，用 zip 打包。该字段仅在复赛阶段提供。
asr	string	苏炳添小组第一苏炳添创造了历史，他成为了第一个进入奥运会百米飞人决战的黄种人。创造了中国田径新的纪录。	视频的音频转文本识别	可能存在空值。
ocr	dict list	[{"time": 0, "text": "苏炳添创造新纪录荣获小组第一"}, ...]	视频的 OCR 识别	该字段为一个列表，记录了不同时刻的 OCR 识别结果。相邻帧的重复识别已被去除。最多提供前 32 秒的 OCR 结果。可能存在空值。

2. 数据集

比赛提供的数据集有三个类别：无标注训练数据集、有标注训练数据集、测试数据集。各类数据集具体包含字段如下表所示。

字段	初赛			复赛		
	训练数据集		测试数据集	训练数据集		测试数据集
	无标注	有标注		无标注	有标注	
id	√	√	√	√	√	√
category_id	×	√	×	×	√	×
title	√	√	√	√	√	√
frames_feature	√	√	√	×	×	×
frames	×	×	×	√	√	√
asr	√	√	√	√	√	√
ocr	√	√	√	√	√	√

3. 提交结果格式

参赛者需要提交所有测试集的 category_id，具体要求如下：

（1） 测试结果写入到一个 csv 文件中进行提交。

（2） csv 文件中包含两列：id 和 category_id，中间用逗号分隔。

（3） csv 文件的行数应与测试集的样本数量相同。视频 id 顺序可以不同。

官方 baseline 代码中 inference.py 有生成提交文件的样例。

二、 评估标准

分类的评估指标采用 F1，由于有多个类别，而且类别不均衡，所以同时采用 F1 micro 和 F1 macro，取平均值。同时，分类体系包含一级分类和二级分类，在评测中会分别计算并取平均值。F1 指标的定义与计算可以参考 sklearn 文档。

最终指标为：

(category1_f1_micro + category1_f1_macro + category2_f1_micro + category2_f1_macro) / 4

考虑实际使用，我们希望参赛选手使用的模型是简单而高效的，不鼓励使用超大模型和各种复杂 ensemble。所以在复赛阶段，我们将限定模型大小并对运行时间做出限制，要求选手提供 docker，包含测试代码，由官方调用。

三、 其他说明

1. 本项比赛全程不允许使用外部数据集；
2. 允许使用开源的词典、embedding 和预训练模型，以上数据和模型需在复赛开始前开源，且需通过邮件的形式报备开源链接地址和 md5，报备邮箱为 wechat_algo@tencent.com。