

全球人工智能技术创新大赛

首届全球人工智能技术创新大赛通知

首届全球人工智能技术创新大赛（Global AI Technology Innovation Contest）是由中国人工智能学会和杭州市余杭区人民政府共同创办，由杭州市未来科技城管委会、阿里云计算有限公司、清华-OPPO 未来终端技术研究中心联合承办，大赛将立足国际视野，聚焦前沿科技与应用创新，推进人工智能领域的学术交流、人才培养、技术发展以及跨界应用与融合，打造人工智能的人才交流平台与产业生态圈。

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。大赛将设置医学影像报告异常检测、PANDA 大场景多对象检测跟踪、小布助手对话短文本语义匹配三个智能算法赛道，欢迎相关领域的学生和研究者积极参与，突破其中的关键性瓶颈，推进下一代人工智能前瞻性研究发展。

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，为在校生成和毕业生参与竞赛提供必要的条件和支持。



首届全球人工智能技术创新大赛

竞赛规程

首届全球人工智能技术创新大赛（以下简称“大赛”）是由中国人工智能学会和杭州市余杭区人民政府共同创办，由杭州未来科技城管委会、阿里云计算有限公司、清华-OPPO 未来终端技术研究中心联合承办，大赛将立足国际视野，聚焦前沿科技与应用创新，推进人工智能领域的学术交流、人才培养、技术发展以及跨界应用与融合，打造人工智能的人才交流平台与产业生态圈。

一、赛道设置

本次大赛将设置医学影像报告异常检测、PANDA 大场景多对象检测跟踪、小布助手对话短文本语义匹配三个智能算法赛道，具体说明详见附件“赛题描述”。

1. 医学影像报告异常检测

影像科医生在工作时会观察医学影像（如 CT、核磁共振影像），并对其作出描述，这些描述中包含了大量医学信息，对医疗 AI 具有重要意义。本赛道的任务要求参赛队伍根据医生对 CT 的影像描述文本数据，判断身体若干目标区域是否有异常以及异常的类型。初赛阶段仅需判断各区域是否有异常；复赛阶段除了判断有异常的区域外，还需判断异常的类型。

2. PANDA 大场景多对象检测跟踪

PANDA (gigaPixel-level humAN centric video DATaset) 是清华大学团队构建的国际上首个动态大场景多对象数据平台，场景平均覆盖平方千米级范围，可同时观测数千人，百米外人脸清晰可识别，视频分辨率近 10 亿像素。本赛道的目的是吸引研究者关注动态大场景多对象数据处理算法的研究，其初赛的任务是行人和车辆多目标检测，即给定静态 PANDA 图像，参赛选手需要检测两种类型的对象：行人和车辆。对于每个行人，需要提交三种检测结果框：可见范围位置框、全身范围位置框、头部范围位置框。对于每辆车，需要提交一个可见范围位置框。复赛的任务是行人多目标跟踪，即给定 PANDA 视频序列，参赛选手需要提交视频中的行人轨迹。

3. 小布助手对话短文本语义匹配

小布助手是 OPPO 公司为欧加集团三品牌手机和 IoT 设备自研的语音助手，为用户提供了有趣、贴心、便捷的对话式服务。意图识别是对话系统中的一个核心任务，而对话短文本语义匹配是意图识别的主流算法方案之一。本赛道要求参赛队伍根据脱敏后的短文本 query-pair，预测它们是否属于同一语义。

二、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，具体组队要求见赛道相关说明；
- 参赛选手应保证报名信息准确有效，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，将禁止以下类型人员报名参赛：

- 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。
- 赞助企业的在职人员（不含实习生）不允许参赛。

三、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，其中初赛均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；复赛阶段测试数据不可见且不可下载，采用 docker 镜像的方式进行提交。决赛要求参赛者进行现场演示和答辩。

1. 报名&组队（1月15日 – 4月7日）

参赛选手须在大赛官网报名并且组队参赛（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。大赛报名系统开放时间为北京时间 2021 年 1 月 15 日 10:00，截止时间为北京时间 2021 年 4 月 7 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；
- 每个选手可以同时选择多个赛道进行参赛，但在每个赛道中只能参加一支队伍，不同赛道可以拥有不同的团队；
- 参赛队伍（包括队长及全体队伍成员）需要在 4 月 7 日中午 12 点前完成实名认证（认证入口：天池官网-右上角个人中心-认证-支付宝实名认证），未完成认证的队伍将无法参加正式比赛。

大赛官方渠道主要包括：

- 大赛官网：<https://gaiic.tianchi.aliyun.com/>
- 大赛邮箱：data@tsinghua.edu.cn
- 大赛 QQ 群：566353409(赛道 1) / 758344321(赛道 2) / 753413531(赛道 3)

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向大赛组委会提交成员更换申请，由参赛队伍全部成员亲笔签名，经由大赛组委会审核后变更生效。

2. 初赛 (3月2日 – 4月9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。若参赛队伍在一天内多次提交结果，新结果版本将覆盖旧版本。

初赛 A 阶段：3月2日 10:00 – 4月7日 20:00，每个参赛队伍每天可以进行 3 次提交，系统立即进行评测并返回成绩。排行榜每小时进行更新，将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 B 阶段：4月8日 10:00 – 20:00，系统将在 4月7日 20:00 更换测试数据，参赛队伍需再次下载加密数据文件，系统将于 4月8日 10:00 公布数据集解密密码。本阶段提供 2 次评测机会，提交截止时间是 4月8日 20:00。排行榜每小时进行更新，将选择参赛队伍在本阶段的历史最优成绩进行排名展示。

初赛结束后，以 B 榜成绩作为比赛算法提交成绩依照，TOP120 团队提交代码审核，要求详见“代码规范”文档，代码提交截止时间是 4月12日 12:00。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，晋级空缺名额将顺延替补。

初赛成绩符合要求的排名前 100 名且通过支付宝实名认证的参赛队伍将晋级复赛，前 10 名队伍还将获得初赛奖励。

3. 复赛 (4月15日 – 5月10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），第一次接触 docker 可直达教程（[链接](#)）。

预热阶段（4月15日 – 4月19日 20:00）：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（4月20日 – 5月10日 12:00）：系统每天提供 3 次测评机会，排行榜每小时进行更新，将选择参赛队伍在本阶段的历史最优成绩进行排名展示。

复赛最终的提交截止时间是 5月10日 12:00，排行榜最后一次刷新时间是 5月11日 12:00。复赛结束后，以正式榜单成绩作为比赛算法提交成绩依照，TOP10 团队提交代码审核，要求详见“代码规范”文档，代码提交截止时间是 5月13日 12:00。

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 (5月中下旬)

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。

四、奖项设置

1. 初赛奖项

对于每个赛道，初赛最终排行榜的前 10 名队伍颁发初赛名次证书。

2. 复赛与决赛奖项

每个赛道的大赛奖金池总额分别为 50 万元人民币，所有奖金均为税前金额。

奖项名称	数量	对象	奖励办法
全国一等奖	3	决赛第 1 名	决赛名次证书，奖金 15 万元
		决赛第 2-3 名	决赛名次证书，奖金 8 万元
全国二等奖	7	决赛第 4-6 名	决赛名次证书，奖金 5 万元
		复赛第 7-10 名	证书，奖金 1 万元
全国三等奖	10	复赛第 11-20 名	证书

3. 周周星

在每个赛道的初赛阶段，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；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。

五、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，所有内容均以大赛官网为准。
3. 本大赛规程的最终解释权归“全球人工智能技术创新大赛”组织委员会所有。

全球人工智能技术创新大赛组织委员会

2021 年 1 月

赛道 1：医学影像报告异常检测

影像科医生在工作时会观察医学影像（如 CT、核磁共振影像），并对其作出描述，这些描述中包含了大量医学信息，对医疗 AI 具有重要意义。本赛道的任务要求参赛队伍根据医生对 CT 的影像描述文本数据，判断身体若干目标区域是否有异常以及异常的类型。初赛阶段仅需判断各区域是否有异常；复赛阶段除了判断有异常的区域外，还需判断异常的类型。判断的结果按照指定评价指标进行评测和排名，得分最优者获胜。

一、比赛数据

1. Sample 数据

医生对若干 CT 的影像描述的明文数据，及描述中有异常区域与异常类型的 label。样本数量为 10 份，以便使参赛队伍对比赛数据有直观的了解（Sample 数据只是为了增进参赛选手对医疗影像描述的直观了解，实际训练与测试数据不一定与 Sample 数据具有相同特征或分布）。

每份样本占一行，使用分隔符“|”分割为 3 列，为不带表头的 CSV 数据格式。

列名	类型	示例
report_ID	int	1
description	string, 影像描述	右下肺野见小结节样影与软组织肿块影
label	由两部分组成。第一部分为若干异常区域 ID, 用空格分割。第二部分为若干异常类型 ID, 用空格分割。两部分用逗号“,”分割。若定义中所有区域均无异常, 则两部分均为空, 此项为“,”。	4,1 2

需要预测的人体区域有 17 个，复赛中需要判断的异常类型有 12 种。由于数据安全需要，不会告知具体区域与类型的名称，只会以 ID 表示，区域 ID 为 0 到 16，类型 ID 为 0 到 11。每个影像描述中可能有零个、一个或多个区域存在异常；若此描述有异常区域，则可能包含一个或多个异常类型。

2. Training 数据

脱敏后的影像描述与对应 label。影像描述以字为单位脱敏，使用空格分割。初赛只进行各区域有无异常的判断，label 只有异常区域 ID。复赛除了判断各区域有无异常，还需要判断各区域异常的类型，因此 label 包含异常区域 ID 与异常类型 ID。初赛 Training 集规模为 10000 例样本，复赛 Training 集规模为 20000 例样本。Training 数据用于参赛选手的模型训练与预估。

初赛 Training 数据格式（不同列使用分隔符“|”分割）：

列名	类型	示例
report_ID	int	1
description	脱敏后的影像描述，以字为单位使用空格分割	101 47 12 66 74 90 0 411 234 79 175
label	由多个异常区域 ID 组成，以空格分隔。若此描述中无异常区域，则为空	3 4

复赛 Training 数据格式（不同列使用分隔符“|”分割）：

列名	类型	示例
report_ID	int	1
description	脱敏后的影像描述，以字为单位使用空格分割	101 47 12 66 74 90 0 411 234 79 175
label	string 由两部分组成：第一部分为若干异常区域 ID，用空格分割；第二部分为若干异常类型 ID，用空格分割。两部分用逗号“，”分割。若定义中所有区域均无异常，则两部分均为空，此项为“，”。	3 4,0 2

3. Test 数据

脱敏后的影像描述，脱敏方法和 Training 相同。Test 数据用于参赛选手的模型评估和排名。初赛 Test 集分为 AB 榜，规模均为 3000。复赛 Test 集规模为 5000。

Test 数据格式（不同列使用分隔符“|”分割）：

列名	类型	示例
report_ID	int	1
description	脱敏后的影像描述，以字为单位使用空格分割	101 47 12 66 74 90 0 411 234 79 175

二、选手提交结果

对于 Test 数据“report_ID,description”，选手应提交“report_ID,prediction”，其中 prediction 是预测结果。初赛中 prediction 是 17 维向量，值在 0 到 1 之间，表示各区域有异常的概率，使用空格分割。复赛中 prediction 是 29 维向量，值在 0 到 1 之间，前 17 个值表示 17 个区域有异常的概率，后 12 个值表示此描述包含各异常类型的概率。

初赛提交数据格式（不同列使用分隔符“|”分割）：

列名	类型	示例
report_ID	int	1
Prediction	17 维向量	0.68 0.82 0.92 0.59 0.71 0.23 0.45 0.36 0.46 0.64 0.92 0.66 0.3 0.5 0.94 0.7 0.38

复赛提交数据格式（不同列使用分隔符“|”分割）：

列名	类型	示例
report_ID	int	1
Prediction	29 维向量（中间不需要使用逗号分隔）	0.68 0.82 0.92 0.59 0.71 0.23 0.45 0.36 0.46 0.64 0.92 0.66 0.3 0.5 0.94 0.7 0.38 0.05 0.97 0.71 0.5 0.64 0.0 0.54 0.5 0.49 0.41 0.06 0.07

三、评估标准

在 Test 数据集上，将对选手提交结果使用 ROC 曲线的 AUC (Area Under Curve) 作为评估标准。在初赛阶段，一个样本对应 17 个预测值，在计算 AUC 时，这 17 个预测值被看作 17 个不同的样

本。即对 N 个测试样本看作 $17N$ 个样本进行 AUC 计算。具体来说，对应某阈值 T ，所有预测值大于 T 的预测即为阳性，反之为阴性。由此得到真阳性 TP、假阳性 FP、真阴性 TN、假阴性 FN，此四个值和为 $17N$ 。对应真阳性率 TPR、假阳性率 FPR 分别为 $TPR = \frac{TP}{TP+FN}$ ， $FPR = \frac{FP}{FP+TN}$ 以 FPR 为横坐标，TPR 为纵坐标，改变阈值 T 即得 ROC 曲线。计算其曲线下面积 AUC 即为最终评价指标。

在复赛阶段，分数由两部分组成。第一部分与初赛相同，对预测值的前 17 维结合真实值计算 AUC，得到 S_1 。第二部分为对所有实际存在异常区域的测试样本，对其预测值后 12 维结合真实异常类型计算 AUC，方法与第一部分相同，若 N 个测试样本中有 M 个实际有异常区域，则将对 $12M$ 个样本进行 AUC 计算（实际无异常的样本不参与第二部分计算），得到 S_2 。复赛最终分数 $S=0.6S_1+0.4S_2$ 。

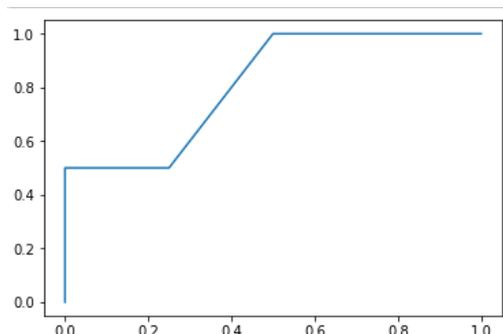
下面以复赛分数为例，举例说明分数计算方法。为了便于说明，这里假设区域有 3 个而不是 17 个，异常类型有 2 个而不是 12 个。设真实值和选手提交的预测值分别如下：

真实值 Ground Truth	预测值 Prediction
0 1 0 1 0	0 0.6 0.7 0.5 0
0 0 1 1 0	0 0.6 0.8 0.1 0.2

则第一部分分数为对以下数据计算 AUC：

真实值	0	1	0	0	0	1
预测值	0	0.6	0.7	0	0.6	0.8

ROC 曲线如下所示，计算其 AUC 得到 $S_1=0.8125$ 。



同理，对类型预测部分的 $2*2=4$ 个样本进行 AUC 计算，得到 $S_2=0.75$ 。因此复赛阶段分数为 $0.6*0.8125+0.4*0.75=0.7875$ 。

四、其他说明

1. 本项比赛全程不允许使用外部数据集。
2. 允许使用预训练模型，如网络模型与 embedding 等。
3. 复赛阶段可以使用初赛阶段的数据集。

赛道 2 : PANDA 大场景多对象检测跟踪

PANDA (gigaPixel-level humAN centric video Dataset) 是清华大学团队构建的国际上首个动态大场景多对象数据平台, 场景平均覆盖平方千米级范围, 可同时观测数千人, 百米外人脸清晰可识别, 视频分辨率近 10 亿像素。本赛道的目的是吸引更多的计算机视觉研究者关注动态大场景多对象数据处理算法的研究, 促使检测、追踪等视觉任务在十亿像素视频数据上得以解决。

本赛道在初赛和复赛两个阶段分别对应不同的赛题:

1. 初赛赛题是行人和车辆多目标检测, 即给定静态 PANDA 图像, 参与者需要检测两种类型的对象: 行人和车辆。对于每个行人, 需要提交三种检测结果框: 可见范围位置框、全身范围位置框、头部范围位置框。对于每辆车, 需要提交一个可见范围位置框。在评价结果时, 会忽略一些特殊情况 (如反射或阴影区域、人与车极为拥挤的区域、严重遮挡区域)。
2. 复赛的赛题是行人多目标跟踪, 即给定 PANDA 视频序列, 参赛选手需要提交视频中的行人轨迹 (每一帧上输出行人 id 和位置框)。

一、 比赛数据

1. Sample 数据

Sample 数据包括 PANDA-Image 和 PANDA-Video 的示例与注释文件, 样本数量各为 1 份, 数据格式与 Training 集格式相同。

数据及标注的可视化如下:

全局视图:



局部视图:



2. Training 数据

2.1 初赛

初赛（PANDA 行人和车辆多目标检测比赛）使用 PANDA-Image 数据集。PANDA-Image 由 555 张静态十亿像素图片组成，总共包含 21 个不同的场景，其中 Training 集包括 390 张图片。Training 集存储在压缩包中，包括每个场景的图片和注释文件（.json 格式）。

Training 集压缩包的目录结构如下所示：

```
|--Train.zip
  |--image_train
    |--scene 1
      |-- XXX.png
      |-- XXX.png
      ...
      |-- XXX.png
    |--scene 2
      |-- XXX.png
      |-- XXX.png
      ...
      |-- XXX.png
    ...
  |--image_annos
    |-- human_bbox_train.json
    |-- vehicle_bbox_train.json
```

human_bbox_train.json 和 vehicle_bbox_train.json 分别包含训练集图像中的行人和车辆的相关注释。数据格式如下：

(1) 人物位置文件（human_bbox_train.json）

人物位置文件中的数据由多个“字典”结构嵌套构成。

字典名称	列名	类型	说明
image_dict	image id	int	图 ID
	image size	image_size	图片大小
	objects list	[object_dict]（格式见下表）	人物字典
image_size	height	int	图片高
	width	int	图片宽

object_dict 中的“category”类一共有“person”，“fake person”，“ignore”，“crowd”四类，“person”表示正常情况的目标，“fake person”，“ignore”，“crowd”分别表示三种特殊情况：反射或阴影区域、严重遮挡区域、人与车极为拥挤的区域，在评价结果时会忽略这三种特殊情况。

“category”为“person”的 object_dict 格式如下：

字典名称	列名	类型	说明
object_dict (person)	category	“person”	人物类别
	pose	"standing" or "walking" or "sitting" or "riding" or "held" (a baby in the arms) or "unsure"	姿态
	riding tipe	"bicycle rider" or "motorcycle rider" "tricycle rider" or "null" (when "pose" is not "riding")	是否骑车
	age	"adult" or "child" or "unsure"	年龄
	rects	[rect_dict1]（格式见下表）	位置

rect_dict1 格式:

字典名称	列名	类型	说明
rect_dict1	head	[rect_dict2] (格式见下表)	头位置
	visible body	[rect_dict2] (格式见下表)	可见位置
	full body	[rect_dict2] (格式见下表)	全身位置

rect_dict2 格式:

rect_dict2	tl	x	float	左上位置
		y	float	
	br	x	float	右下位置
		y	float	

"x", "y": 0 到 1 之间的浮点数, 分别表示坐标与图像的宽度和高度的比值。

“category” 为 “fake person”, “ignore” 或 “crowd” 的 object_dict 格式如下:

字典名称	列名	类型	说明
object_dict (ignore/fake/crowd)	category	"ignore" (someone who is heavily occluded) or "fake person" or "crowd" (extremely dense crowd)	人物类别
	rect	[rect_dict2] (格式见下表)	位置

rect_dict2 的格式如下:

rect_dict2	tl	x	float	左上位置
		y	float	
	br	x	float	右下位置
		y	float	

"x", "y": 0 到 1 之间的浮点数, 分别表示坐标与图像的宽度和高度的比值。

(2) 车辆位置文件 (vehicle_bbox_train.json)

车辆位置文件中的数据由多个“字典”结构嵌套构成。

字典名称	列名	类型	说明
image_dict	image id	int	图 ID
	image size	image_size	图片大小
	objects list	[object_dict] (格式见下表)	车辆字典
image_size	height	int	图片高
	width	int	图片宽

object_dict 格式:

字典名称	列名	类型	说明
object_dict	category	"small car" or "midsize car" or "large car" or "bicycle" or "motorcycle" or "tricycle" or "electric car" or "baby carriage" or "vehicles" or "unsure"	目标类别
	rect	rect_dict2	位置

"vehicles"是指密集的车辆群，应忽略；

"small car", "midsize car", "large car"是四轮以上的机动车辆，以车辆大小来区分；

"electric car"是电动观光车、电动巡逻车；

"x", "y"是 0 到 1 之间的浮点数，分别表示坐标与图像的宽度和高度的比值。

“rect_dict2”是格式与人物位置文件中“rect_dict2”的格式相同。

2.2 复赛

复赛（PANDA 行人多目标跟踪比赛）使用 PANDA-Video 数据集。PANDA-Video 数据集由 15 段视频序列组成。由于数据规模的限制，PANDA-Video 经过了抽帧处理，抽帧后的帧率为 2FPS，并以帧图像的方式存储视频。

其中 Training 集包含 10 段视频序列，Training 集压缩包的目录结构如下所示：

```
|--Train.zip
  |--video_train
    |--scene 1
      |-- XXX.png
      |-- XXX.png
      ...
      |-- XXX.png
    |--scene 2
      |-- XXX.png
      |-- XXX.png
      ...
      |-- XXX.png
    ...
  |--video_annos
    |--scene 1
      |-- tracks.json
      |-- seqinfo.json
    |--scene 2
      |-- tracks.json
      |-- seqinfo.json
    ...
```

tracks.json 包含行人轨迹标注信息，seqinfo.json 包含视频序列信息，数据格式如下：

行人轨迹标注文件（tracks.json）

字典名称	列名	类型	说明
track_dict	track id	int	行人 ID
	frames	[frame_dict]（格式见下表）	行人信息

frame_dict 格式：

字典名称	列名	类型	说明
frame_dict	frame id	int	帧 ID
	rect	[rect_dict]（格式见下表）	行人位置
	face orientation	"back" or "front" or "left" or "left back" or "left front" or "right" or "right back" or "right front" or "unsure"	脸朝向
	occlusion	"normal" or "hide" or "serious hide" or "disappear"	是否是遮挡

rect_dict 格式:

rect_dict	tl	x	float	左上角坐标
		y	float	
	br	x	float	右下角坐标
		y	float	

视频序列文件 (seqinfo.json)

列名	类型	说明
name	scene_name	视频名称
frameRate	int	帧率
seqLength	int	视频长度
imWidth	int	宽
imHeight	int	高
imExt	file_extension	图片格式
imUrls	image_url	图片位置相对路径

3. Test 数据

3.1 初赛

Test 集包括 165 张图片，组织格式与 Training 集基本相同。Test 集存储在压缩包中，包括每个场景的图片和注释文件 (.json 格式)。

Test 集压缩包的目录结构如下所示:

```
|--Test.zip
  |--image_train
    |--scene 1
      |-- XXX.png
      |-- XXX.png
      ...
      |-- XXX.png
    |--scene 2
      |-- XXX.png
      |-- XXX.png
      ...
      |-- XXX.png
    ...
  |--image_annos
    |-- human_bbox_test.json
    |-- vehicle_bbox_test.json
```

human_bbox_test.json 和 vehicle_bbox_test.json 只包含测试集图像的图像路径、图像 id 以及图像大小。数据格式如下:

人物位置文件 (human_bbox_test.json)

字典名称	列名	类型	说明
image_dict	image id	int	图 ID
	image size	image_size	图片大小

车辆位置文件 (human_bbox_test.json)

字典名称	列名	类型	说明
image_dict	image id	int	图 ID
	image size	image_size	图片大小

3.2 复赛

Test 集包含 5 段视频序列，组织格式与 Training 集基本相同。

Test 集压缩包的目录结构如下所示：

```
|--Test.zip
  |--video_test
    |--scene 1
      |-- XXX.png
      |-- XXX.png
      ...
      |-- XXX.png
    |--scene 2
      |-- XXX.png
      |-- XXX.png
      ...
      |-- XXX.png
    ...
  |--video_annos
    |--scene 1
      |-- seqinfo.json
    |--scene 2
      |-- seqinfo.json
    ...
```

Test 集只包含 seqinfo.json，数据格式与 Training 集相同。

二、选手提交结果

1. 初赛

提交的结果文件的格式与 [COCO Challenge](#) 的格式类似。我们要求参与者以单个 det_results.json 文件（通过 Matlab 的 gason 保存或 Python 的 json.dump 保存）提交结果。det_results.json 文件应该包含一个列表，其中的每个元素都是一个字典。每个字典包含检测结果的信息，具体文件示例如下：

```
[
  {
    "image_id": 1,
    "category_id": 0,
    "bbox_left": 0.63,
    "bbox_top": 0.37,
    "bbox_width": 20,
    "bbox_height": 30,
    "score": 0.8
  },
  {
    "image_id": 1,
    "category_id": 1,
    "bbox_left": 0.53,
    "bbox_top": 0.27,
    "bbox_width": 40,
    "bbox_height": 50,
    "score": 0.7
  }
  ...
]
```

每个字典各字段含义如下：

名称	说明
image_id	图像序列号，应与标注文件一致
category_id	检测结果的类型框,应与下表一致
bbox_left	预测的边界框左上角的 x 坐标
bbox_top	预测对象边框左上角的 y 坐标
bbox_width	预测对象边界框的像素宽度
bbox_height	预测对象包围框的像素高度
score	包围对象实例的预测边框的置信度，[0,1]

目标	category_id
人体可见部分	1
全人体	2
人头	3
车辆可见部分	4

注意，对于提交的测试结果，图像 id 应该与注释文件中的图像 id 相同，

2. 复赛

提交的结果文件的格式与 [MOTChallenge](#) 的格式相同。我们要求参与者以单个 mot_results 文件夹提交结果。mot_results 文件夹中的每个.txt 文件表示每个序列的结果。文件名必须与序列名称完全相同(例如 01_University_Canteen.txt, 区分大小写)。每个.txt 文件的格式与 MOTChallenge 的格式相同，每行表示一个目标某一帧的跟踪结果。每一行必须包含 10 个值，最后 4 项固定为-1。

提交的 scene.txt 文件示例：

```
1 1 0.63 0.37 20 30 -1 -1 -1 -1
1 2 0.53 0.29 15 29 -1 -1 -1 -1
1 3 0.89 0.66 38 62 -1 -1 -1 -1
...
```

每一行前 6 项各值含义如下：

位置	名称	描述
1	frame	帧的 ID
2	id	目标的 ID
3	bb_left	预测的边界框左上角的 x 坐标
4	bb_top	预测对象边框左上角的 y 坐标
5	bb_width	预测对象边界框的像素宽度
6	bb_height	预测对象包围框的像素高度

三、评估标准

1. 初赛

为了评估检测算法效果，我们使用 AP, AP_{IOU=0.50}, AP_{IOU=0.75}, AR_{max=10}, AR_{max=100}, AR_{max=500} 指标。初赛得分依据于 AP 和 AR_{max=500} 两项指标的调和平均数，排名完全依据于初赛得分，高者为优，计算公式如下：

$$\text{Score1} = (2 * \text{AP} * \text{AR}_{\text{max}=500}) / (\text{AP} + \text{AR}_{\text{max}=500})$$

AP 和 AR 的计算方式同 [COCO 数据集](#)，参考公式如下：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\# \text{ ground truths}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\# \text{ predictions}}$$

$$\text{AP}[\text{class}] = \frac{1}{\# \text{ thresholds}} \sum_{\text{iou} \in \text{ thresholds}} \text{AP}[\text{class}, \text{iou}]$$

$$\text{AP} = \frac{1}{\# \text{ classes}} \sum_{\text{class} \in \text{ classes}} \text{AP}[\text{class}]$$

$$m\text{AP}_{\text{COCO}} = \frac{m\text{AP}_{0.50} + m\text{AP}_{0.55} + \dots + m\text{AP}_{0.95}}{10}$$

相关评估代码可以在 [PANDA-Toolkit](#) 中找到。

2. 复赛

为了评价多目标跟踪算法的性能，我们使用 MOTA 和 MOTP 指标。复赛得分依据于 MOTA 和 MOTP 两项指标的调和平均数，高者为优。排名依据初赛与复赛得分的加权得分，计算公式如下：

$$\text{Score2} = (2 * \text{MOTA} * \text{MOTP}) / (\text{MOTA} + \text{MOTP})$$

$$\text{Final Score} = w * \text{Score1} + \text{Score2}$$

w 为初赛分数所占的权重，我们设置 w=0.2。

MOTA 和 MOTP 的计算方式同 [COCO 数据集](#)，参考公式如下：

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + f_{p_t} + mme_t)}{\sum_t g_t}$$

$$\text{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}$$

评估代码可以在 [PANDA-Toolkit](#) 中找到。

四、其他说明

1. 初赛阶段和复赛阶段若需要使用外部的数据集和预训练模型，仅允许使用 ImageNet/COCO 和在 ImageNet/COCO 上的预训练模型。
2. 复赛阶段允许使用初赛阶段的数据集。
3. 提交的模型的大小（Param Mem）必须小于 2G 且模型的 FLOPs 小于 3×10^{10} flops，初赛和复赛单张图片处理时间不超过 30 秒，否则视为无效。

赛道 3：小布助手对话短文本语义匹配

小布助手是 OPPO 公司为欧加集团三品牌手机和 IoT 设备自研的语音助手，为用户提供了有趣、贴心、便捷的对话式服务。意图识别是对话系统中的一个核心任务，而对话短文本语义匹配是意图识别的主流算法方案之一。本赛道要求参赛队伍根据脱敏后的短文本 query-pair，预测它们是否属于同一语义，提交的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。

一、比赛数据

1. 训练数据：训练数据包含输入 query-pair，以及对应的真值。初赛训练样本 10 万，复赛训练样本 30 万，这份数据主要用于参赛队伍训练模型，为确保数据的高质量，每一个样本的真值都有进行人工标注校验。每行为一个训练样本，由 query-pair 和真值组成，每行格式如下：

- query-pair 格式：query 以中文为主，中间可能带有少量英文单词（如英文缩写、品牌词、设备型号等），采用 UTF-8 编码，未分词，两个 query 之间使用\t 分割。
- 真值：真值可为 0 或 1，其中 1 代表 query-pair 语义相匹配，0 则代表不匹配，真值与 query-pair 之间也用\t 分割。

训练数据样本举例（空白间隔为\t）：

```
肖战的粉丝叫什么名字    肖战的粉丝叫什么    1
王者荣耀里面打野谁最厉害    王者荣耀什么英雄最好玩    0
我想换个手机    我要换手机    1
我是张睿    我想张睿    0
不想    不想说    0
```

2. 初赛与复赛测试样本：脱敏后的 query-pair 数据，初赛采用 A/B 榜的方式，A 榜和 B 榜样本规模分别为 5 万，复赛样本规模 5 万（与初赛不重复）。

测试数据样本举例（空白间隔为\t）：

```
肖战的粉丝叫什么名字    肖战的粉丝叫什么
王者荣耀里面打野谁最厉害    王者荣耀什么英雄最好玩
我想换个手机    我要换手机
我是张睿    我想张睿
不想    不想说
```

二、提交结果

选手针对测试数据提交预测结果文件，结果文件中每行为一个预测值，是 0 与 1 之间的一个浮点数，代表 query-pair 语义匹配的概率，与测试数据每行一一对应。

结果文件内容举例：

```
0.001
0.999
```

三、评估标准

比赛的评估标准由性能标准和效果标准两部分组成，初赛采用效果标准，AUC 指标，具体定义如下：

$$AUC = \frac{\sum_{i \in \text{语义匹配样本集}} \text{rank}(i) - \frac{M(1+M)}{2}}{M * N}$$

其中：

rank(i)：表示 i 这个样本的预测得分在测试集中的排序；

M：测试集中语义匹配的样本的个数；

N：测试集中语义不匹配的样本的个数。

复赛阶段将同时考虑性能标准和效果标准，效果标准继续采用 AUC 指标，性能标准是约束条件，在复赛阶段需要在限定时间内完成预测，具体要求详见复赛开始前提供的“容器镜像”页面。

四、其他说明

1. 本项比赛全程不允许使用外部数据集。
2. 允许使用预训练模型，如网络模型与 embedding 等。
3. 复赛阶段允许使用初赛阶段的数据集。