# 第一届全国信息隐藏大赛通知

随着网络信息技术的迅猛发展，在各国政府推动下，网络安全的重要性正日益提升。随着大数据技术、人工智能等革命性技术的发展，网络安全早已不再是一个孤立的学科问题，也不仅仅是某项单一技术成果就能解决的重大安全问题。信息隐藏技术作为一种维护网络空间安全的重要技术手段，在核心数据保护、网络安全通信、数字版权保护、人工智能安全等领域发挥了重要的作用。为了进一步推动网络空间信息隐藏技术的发展，中国电子学会通信分会、北京电子技术应用研究所和清华大学联合举办第一届全国信息隐藏大赛（The 1st Chinese Information Hiding Competition，简称 CIHC 2019）。本大赛以促进信息隐藏的学科发展和成果转换为目的，依托第十五届全国信息隐藏暨多媒体信息安全学术大会（CIHW2019），面向社会各界开放。竞赛过程保证公平公正，符合相关法律法规。本次比赛相关安排和规则如下：

## 【赛程】

本次大赛分为图像、音频、文本三个项目分别进行，分为线上竞赛和线下颁奖两个部分，具体安排和要求如下：

- 线上竞赛（2019 年 6 月 15 日——2019 年 10 月 1 日）：
  - 报名成功的各队队员按照竞赛规则，从网站下载数据集（数据集地址：https://drive.google.com/open?id=1i2jHdhI19Ge6lBe36am-cfFNYjwih9lL），本地调试模型，并在规定时间内在线提交检测结果；
- 线下颁奖（2019 年 10 月 19 日——2019 年 10 月 20 日）：
  - 根据初赛中各项目参赛队伍最终成绩，选取各组排名前 3 名队伍进入线下颁奖环节；
  - 进入该环节的队伍将在第十五届全国信息隐藏暨多媒体信息安全学术大会（CIHW 2019）上接受颁奖。

## 【报名】

1，参赛队伍可自行命名队伍名字，名字可包括简体汉字、字母、数字以及通用中英文标点字符，但不得包含违法违规、歧视性、不良信息等内容。不同参赛队伍名称相同时，通过自行协商解决。

2，每支参赛队伍人数不得超过三人。

3，每支参赛队伍需填写第一届全国信息隐藏大赛报名表，报名表下载链接（http://www.cihw.org.cn/index.php/9/），需要提供所在单位名称、参赛人姓名、身份证号和联系方式等信息。填写完报名表后发送到组委会邮箱（**CIHC2019@163.com**），未提交参赛队伍信息表或信息不全视为未参赛。

4，参赛队伍在提交报名表后如需添加成员、变更所在单位等信息，需要联系组委会说明情况，并重新提交参赛队伍信息表。

5，参赛成员可向组委会邮箱提交退出声明退出参赛队伍。如所有成员退出，该队伍默认为自行解散。

5，选手需确保报名信息准确有效，组委会有权取消不符合条件队伍的参赛资格及奖励。
6，参赛队伍不得以任何形式通过不正当手段干扰比赛进行、作弊等。
7，报名阶段需选择报名项目：图像，音频，文本。每个项目下，每位参赛人员只能加入一个参赛队伍。允许参赛人员跨项目组加入其他项目的队伍。

## 【规则】

1，报名成功后，参赛队伍根据分组可以下载比赛数据，本地调试算法，然后在规定时间内（2019 年 6 月 15 日——2019 年 10 月 1 日）在线提交结果，提交地址为 http://www.cihw.org.cn/ 网站上的**【比赛提交】**栏目（提交通道于 2019 年 6 月 17 日开放）；
2，参赛队伍在截止前可多次提交结果文件，最终成绩以最后一次提交的文件为准；
3，参赛队伍提交的结果文件不符合格式要求视同为无效；
4，比赛结果提交的截止时间为 2019 年 10 月 1 日 23:59，超过该时间提交的结果无效；
4，本次竞赛结果以 Accuracy 为得分按照从高到低进行排名判定。得分相同的情况下，参考 F1-score；
5，比赛过程中，会不定期随机抽检 50%测试样本计算得分，并公布临时排名，临时排名公布地址：(http://www.cihw.org.cn/index.php/9/)。最终排名以完整测试集的检测得分为依据；
6，若提交的文件无法解码或者提交最终结果格式不对，将导致没有成绩；

## 【数据说明】

### ■ 图像隐写检测数据集

- **下载路径**：【https://drive.google.com/open?id=144ImZ28zmvN_dZbbeJTyLjz8u8nGklSY】
- **测试图像数量**：1 万张
- **测试图像来源**：源自不同手机型号拍摄的原始未经处理的照片；
- **测试图像尺寸**：原始照片按最短边裁剪成正方形，然后统一缩放成 1024*1024 大小；
- **测试图像品质因子**：筛选出品质因子分布在 90-95 之间的图像；
- **测试隐写算法**：以 0.5 的概率选择数据集中的图像进行嵌入，嵌入算法选择三种不同的隐写算法以不同的嵌入率进行随机嵌入，三种隐写算法及相应的论文如下：

| 隐写算法 | 相关论文 |
|---|---|
| **J_unward** | V. Holub, J. Fridrich, T. Denemark, Universal Distortion Function for Steganography in an Arbitrary Domain, EURASIP Journal on Information Security |
| **nsF5** | J. Fridrich, T. Pevný, and J. Kodovský, Statistically undetectable JPEG steganography: Dead ends, challenges, and opportunities. |
| **UERD** | Using Statistical Image Model for JPEG Steganography: Uniform Embedding Revisited |

- **辅助训练集和验证集**：

为了便于参赛队伍训练和调试模型，我们提供一个辅助训练集和验证集。该辅助训

练集包含 10 万张来自网站 Unsplash（https://unsplash.com）的原始图像。经过和测试集图像同样的裁剪、缩放操作变成 1024*1024 大小，以和测试集同样的隐写策略进行随机比特流嵌入，最终获得 10 万对 cover-stego 图像（路径：./Image/train）。

*注意：允许各参赛队使用额外数据对参赛模型进行训练和调试。

考虑到训练集和测试集图像来源不一致，因此我们从测试集中随机抽取 1809 张图像构成验证集（路径：./Image/valid），并提供验证集标签（路径：./Image/valid_labels.txt）。参赛队伍可以根据验证集测试结果调试模型。

## ■ 音频隐写检测数据集

- **下载路径：**【https://drive.google.com/open?id=1bjqlB4QjoLlB6ylvAQJ3ZjvU6lRX8Lpk】
- **音频来源**：采集不同人的语音信号，然后按照 G.729a 标准进行编码；
  ***提示：****不熟悉该编码方式，可以参考这篇论文的建模和处理方式：*
  *Real-Time Steganalysis for Stream Media Based on Multi-channel Convolutional Sliding Windows*
- **音频时长**：编码后的语音信号统一裁剪成 1s 的语音片段；
- **隐写算法：**每段语音样本随机选择两种不同的隐写算法之一以随机的嵌入率进行嵌入，两种隐写算法名称及相应的论文如下：

| 隐写算法 | 相关论文 |
|---|---|
| CNV-QIM | Xiao B, Huang Y, Tang S. An approach to information hiding in low bit-rate speech stream |
| Pitch | Huang Y, Liu C, Tang S, et al. Steganography integration into a low-bit rate speech codec |

- **语音隐写分析提供训练数据和测试：**

- **训练集**：

音频数量：隐写和非隐写音频各 155327 段。隐写和非隐写音频由相同的 wav 音频（约 43 个小时的中文音频）编码得到。

文件夹：

./ch_0_g729a：没有隐写的 G.729a 编码后音频

./ch_steg_g729a：有隐写的 G.729a 编码后音频。

- **测试集：**

2000 段编码后时长为 1s 的音频片段，以 0.5 的概率选择数据集中的音频片段，然后随机挑选两种嵌入算法之一，并将随机生成的比特流以随机的嵌入率进行随机嵌入。

## ■ 文本隐写检测数据集

- **下载路径：**【https://drive.google.com/open?id=17n4VjK_B3nfmv7ppgmf32Tbflb6t64pl】
- **文本数量：**240,000 句子
- **文本语言：**英文

- **文本来源：**原始文本通过自动爬取多个社交网站上的文本，经过去除广告链接，去除特殊字符等预处理后构成。
- **隐写算法：**我们采用四种隐写算法将随机生成的比特流嵌入文本中，其中两种文本修改式隐写算法和两种文本自动生成式隐写算法，其隐写算法名称和对应的论文如下所示：

| 隐写算法 | 相关论文 |
|---|---|
| 修改式 | Lexical steganography through adaptive modulation of the word choice hash |
| | Adaptive Text Steganography by Exploring Statistical and Linguistical Distortion |
| 生成式 | Yang Z, Jin S, Huang Y, et al. Automatically Generate Steganographic Text Based on Markov Model and Huffman Coding |
| | Yang Z L, Guo X Q, Chen Z M, et al. RNN-stega: Linguistic steganography based on recurrent neural networks |

## 【提交】

- 保存格式：

1）若预测该样本为隐写数据，则预测标签为 1；若预测该样本为非隐写数据，则预测标签为 0。

2）按照测试数据标号，每一行写入一个预测标签。例如对于图像组，第一行为 1.jpg 的预测标签，第二行为 2.jpg 的预测标签。

3）图像保存成【队名+Image_label.txt】，音频保存成【队名+Voice_label.txt】，文本保存成【队名+Text_label.txt】。

例如：

0
1
0
1
0
1

- 将预测结果在比赛规定时间内（2019 年 6 月 15 日——2019 年 10 月 1 日）提交，提交地址为 http://www.cihw.org.cn/ 网站上的**【比赛提交】**栏目（提交通道于 2019 年 6 月 17 日开放）。

## 【排名】

本赛题评分以 accuracy 为准，得分相同时，参照 F1-score 排序。选手预测结果和真实标签进行比对，几个数值的定义先明确一下：

True Positive（TP）表示正确识别的隐写样本的个数;
False Positive（FP）表示错误的隐写判定的样本个数;
True Negative（TN）数值表示正确识别的非隐写的样本个数;
False Negative（FN）数值表示错误判定为非隐写的样本个数。

基于此，我们就可以计算出 precision , recall , accuracy 和 F1-score：

$$precision = TP / (TP + FP)$$
$$recall = TP / (TP + FN)$$
$$accuracy = (TP + TN) / (TP + FP + TN + FN)$$
$$F1\text{-}score = 2 * precision * recall / (precision + recall)$$

**【奖励】**

各项目的比赛获奖团队将根据名次获得证书及相应的奖励。

**【联系方式】**


清华大学 杨忠良 先生

邮箱： CIHC2019@163.com
手机：18811536956