

# 面向图像描述生成自动评测方法的共享任务

## 1. 任务背景

图像描述生成是自然语言生成研究的一个重要方向，这项任务涉及图像理解和文本生成两个方面的研究问题，即如何自动生成自然语言来描述图像表达的主要语义。图像描述生成涉及到自然语言理解和计算机视觉领域两个领域的多个方面的问题，近年来受到学术界和工业界的广泛关注。

随着深度学习技术的发展，端到端图像描述生成系统的表现有了大幅度提升。然而，如何评价图像描述生成系统一直是困扰各界的难题。人工评测的结果具有较强的信度和效度，但存在过程费时费力的问题。因此许多的图像描述生成系统采用自动评测方法，最常用评测指标包括 BLEU, METEOR 等。但自动评测方法也存在许多不足之处，如强调生成文本与标准答案之间的 n-gram 重叠，而不考虑生成文本的准确性来评价系统。

为了深入理解并解决图像描述生成系统评测问题，先提出本共享任务：**我们邀请参与者针对图像描述生成系统的自动评测方法展开研究，并提交其研发的自动评测算法。我们希望参与者得到的评测结果尽可能地与人工评测的结果一致。**通过本共享任务，我们希望能够吸引更多研究者和开发者关注文本生成的评测问题，通过针对评测方法的研究进一步推动自然语言生成理论和实践问题的研究水平。

本届评测得到中国中文信息学会自然语言生成专业委员会（筹）支持，将在2021年11月20日首届中国自然语言生成大会(CCNLG-2021)召开评测研讨会，并在大会上对获奖团队颁奖。

## 2. 任务描述

本共享任务邀请参与者提交一个或多个面向图像描述生成任务的自动评测算法，算法的目标是使自动评测方法给出的分数与人工评测的分数尽可能一致。我们将为参与者提供研究所需数据集，并采用客观的评价指标结果作为提交算法的最终成绩。

**任务描述：**提出面向图像描述生成任务的评测方法，利用该方法对自动生成的图像描述进行打分，并使自动评测结果尽量接近于人工评测结果。

**数据集（数据截取自公开数据集）：**数据集包括以下三部分内容：


- (1) 自动生成的图像描述文本集合；
- (2) 人工撰写的图像描述文本集合；
- (3) 文本集（1）的人工评分集；

**提交结果：**请参与者提交以下材料：

- (1) 提出的自动评测算法的相关技术文档；
- (2) 提出的自动评测算法在我们提供的测试集上的评分结果。

**评价指标：**我们将通过衡量参与者提出的自动评测方法与人工评测方法之间的相关性来评价自动评测算法的好坏，评价指标将采用 Kendall 协调系数（Kendall Correlation）。

**数据集实例：**

	自动生成文本	一个女人站在一群人的背景中。
	人工撰写文本	一群人走在繁忙的人行道上。
	人工评分结果	4.0

### 3. 时间安排

	时间	说明
报名注册	10月10日至10月20日	发布共享任务说明，接受参与者报名
发布训练集	10月21日	发布训练数据集
发布测试集	11月9日	发布测试数据集
提交结果	11月10日	提交测试数据集的结果和技术报告
评测会议	11月20日	发布比赛最终结果，进行会议

### 4. 发起单位

发起单位：青海师范大学、中央民族大学

指导单位：中国中文信息学会自然语言生成专委会（筹）



评测委员会:

李琳, 青海师范大学

赵小兵, 中央民族大学

联系方式: [meetingqhnu@sohu.com](mailto:meetingqhnu@sohu.com)

## 5. 反作弊声明

(1) 参与者禁止注册多账户报名, 经发现将取消成绩并严肃处理。

(2) 参与者禁止在指定考核技术能力的范围外利用规则漏洞或技术漏洞等不良途径提高成绩排名, 经发现将取消成绩并严肃处理。

(3) 可以接触到赛题相关数据的人员, 其提交结果将不计入排行榜及评奖。

## 6. 交流平台

主办方建立技术讨论群(微信群), 供选手讨论、沟通, 主办方也将安排工作人员定期在群内答疑, 且后续的相关活动信息均会在群内发布。

## 参考文献

1. S Aditya, Yang, Y., C Baral, Fermuller, C. , & Aloimonos, Y.. (2015). From images to sentences through scene description graphs using commonsense reasoning and knowledge. *Computer Science*, 1549-1553.
2. Li, X., Xu, C., Wang, X., Lan, W., Jia, Z., & Yang, G., et al. (2018). Coco-cn for cross-lingual image tagging, captioning and retrieval. *IEEE Transactions on Multimedia*.
3. Hodosh, M., Young, P., & Hockenmaier, J.. (2015). Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47(1), 853-899.
4. Li, X., Lan, W., Dong, J., & Liu, H.. (2016). Adding Chinese Captions to Images. *Acm on International Conference on Multimedia Retrieval*. ACM.
5. Cui, Y., Yang, G., Veit, A., Huang, X., & Belongie, S.. (2018). Learning to

Evaluate Image Captioning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE.

6. Lee, H., Yoon, S., Derroncourt, F., Kim, D. S. , & Jung, K.. (2020). ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems.