CCL24-Eval任务9总结报告：中文图文多模态理解评测

王宇轩[1]，刘议骏[2]，万志国[1]，车万翔[2]
[1]之江实验室，杭州，311121
[2]哈尔滨工业大学，哈尔滨，150001
yxwang@zhejianglab.com, yijunliu@ir.hit.edu.cn
wanzhiguo@zhejianglab.com, car@ir.hit.edu.cn

摘要

中文图文多模态理解评测任务旨在从多角度评价中文图文多模态预训练模型的图文多模态建模和理解能力。本任务共包括五个子任务：图片检索、文本检索、视觉问答、视觉定位和视觉对话，最终成绩根据这五个任务的得分综合计算。本文首先介绍了任务的背景和动机，然后从任务介绍、评价指标、比赛结果、参赛方法等方面介绍并展示了本次评测任务的相关信息。本次任务共有11支队伍报名参赛，其中3支队伍提交了结果。

关键词： 中文；多模态；图文检索；视觉问答；视觉定位；视觉对话

Overview of CCL24-Eval Task 9: Chinese Vision-Language Understanding

Yuxuan Wang[1], Yijun Liu[2], Zhiguo Wan[1], Wanxiang Che[2]
[1]Zhejiang Lab, Hangzhou, 311121
[2]Harbin Institute of Technology, Harbin, 150001
yxwang@zhejianglab.com, yijunliu@ir.hit.edu.cn
wanzhiguo@zhejianglab.com, car@ir.hit.edu.cn

Abstract

The Chinese Vision-Language Understanding Task aims to evaluate the vision-language modeling and understanding capabilities of Chinese vision-language pre-training models from multiple perspectives. This task includes five sub-tasks: image retrieval, text retrieval, visual question answering, visual grounding, and visual dialogue. The final score is calculated based on the combined results of these five tasks. This paper first introduces the background and motivation of the task, then presents and demonstrates relevant information about this task from aspects including task description, evaluation metrics, submitted results, and participant methods. A total of 11 teams registered to participate in this task. And 3 teams eventually submitted their results.

Keywords: Chinese , Multimodality , Image-text retrieval , Visual question answering , Visual grounding , Visual dialog

1 背景和动机

近年来，英文图文数据集经历了快速发展，从最基本的图像描述任务开始。继该领域的MS-COCO (Lin et al., 2014)和Flickr30K (Young et al., 2014)图片描述数据集之后，大量涵盖各种任务的英文图文数据集相继出现，这些任务包括视觉问答 (Antol et al., 2015; Goyal

第二十三届中国计算语言学大会论文集，第364页–第371页，太原，中国，2024年7月25日至28日。
卷3：评测报告

et al., 2017)、视觉推理 (Suhr et al., 2017; Suhr et al., 2019; Zellers et al., 2019)、视觉定位 (Kazemzadeh et al., 2014; Mao et al., 2016)、视觉蕴涵 (Xie et al., 2019)、视觉关系检测 (Plummer et al., 2015)和视觉对话visual dialogue (Das et al., 2017)等。这些英文图文数据集的出现对英文图文多模态预训练模型的评价体系建立起到了重要作用，也同时推动了该领域的发展。

在其他语言上，近年来也有不少工作尝试在这些数据集的基础上构建非英语的图文数据集。例如，MS-COCO数据集就被扩展到了德语、法语 (Rajendran et al., 2016)、日语 (Yoshikawa et al., 2017)和中文 (Li et al., 2019)上。这些语言上的图文数据集或者是直接将原始标注中的英语翻译成目标语言，或者是使用目标语言在MS-COCO的图片上重新进行标注。然而，无论通过上述哪一种方法构建的数据集，都使用了原始英文图文数据集中来自西方文化背景的图片。研究表明，使用这种包含文化偏置的数据会严重限制模型在很多其他语言和文化中的表现。(Stock and Cissé, 2018; DeVries et al., 2019; Liu et al., 2021) 同理，在这类包含文化偏置的图片基础上构建的中文图文数据集也无法客观准确地评价目前的中文图文预训练模型。

针对该问题，我们组织了本次中文图文多模态理解评测任务，从收集图片流程开始，严格控制图片为中国文化环境中具有代表性或日常生活常见的内容，并选择了5个重要且具有代表性的图文多模态理解任务：

- 图片检索（Image Retrieval）：基于给定的文本描述从若干候选中检索出对应图片。
- 文本检索（Text Retrieval）：基于给定的图片从若干候选中检索出对应的文本描述。
- 视觉问答（Visual Question Answering）：基于给定的图片用短语回答问题。
- 视觉定位（Visual Grounding）：基于给定的图片和文本描述找出图片中对应的实体。
- 视觉对话（Visual Dialog）：基于给定的图片和对话历史从若干候选中选出最合适的回复文本。

本次评测旨在通过上述5个子任务从图文表示对齐、图文理解和推理、图片细节理解和图片整体理解等多个角度对中文图文预训练模型进行评价。

2 评测任务及数据

2.1 图片类别及图片收集

本任务包含15大类、92小类的图片，具体图片类别如表 1所示。

保证数据集中的图片具有中文文化代表性，是本次评测在构建数据集时十分重要的一点。我们使用百度众包从互联网上采集图片，并且在培训过程中着重强调了收集的图片内容为中国文化环境中具有代表性或日常生活常见的。此外，为了保证后续任务具有足够挑战性，我们对每个类别的图片都采集了两个不同的子集。其中第一个子集中的图片必须包含该图片类别的至少2个实体，该子集用于后续视觉问答和视觉定位任务的标注，这是为了保证在视觉定位任务的标注过程中能让模型区分同一类别的不同实体，而非更简单的区分不同类别的实体。而第二个子集中的图片必须包含3到5个不同类别的实体，该子集用于后续视觉对话任务的标注，这是为了保证在视觉对话任务的标注中对话内容足够丰富。此外，这两个子集上都会进行图文检索任务的标注。

为了进一步确定收集到的图片的质量，我们对通过外包收集来的图片进行了二次筛查，以确保这些图片都具有中文文化代表性，并且各子类中的图片都符合对应子类的要求。通过二次筛查，我们过滤掉了大部分不符合要求的图片。为了进一步确保标注质量，在后续标注过程中，当标注人员发现图片不符合上述要求时，我们也允许标注人员跳过当前图片。

2.2 图文检索

图文检索包括两个子任务，分别是图片检索和文本检索，其中图片检索任务定义为给定一句文本，要求模型从候选图片集合中选出最相关的10张图片，并对它们进行排序。而文本检索任务定义为给定一张图片，要求模型从候选文本集合中选出最相关的10个文本，并对它们进行

第二十三届中国计算语言学大会论文集，第364页-第371页，太原，中国，2024年7月25日至28日。
卷3：评测报告

大类	图片类别
动物	大熊猫, 牛, 鱼, 狗, 马, 鸡, 鼠, 鸟, 人, 猫
食物	火锅, 米饭, 饺子, 面条, 包子
饮品	奶茶, 可乐, 牛奶, 茶, 粥, 酒
衣服	汉服, 唐装, 旗袍, 西装, T恤
植物	柳树, 银杏, 梧桐, 白桦, 松树, 菊花, 牡丹, 兰科, 莲, 百合
水果	荔枝, 山楂, 苹果, 哈密瓜, 龙眼
蔬菜	小白菜, 马铃薯, 大白菜, 胡萝卜, 花椰菜
农业	锄头, 犁, 耙, 镰刀, 担杖
工具	汤勺, 碗, 砧板, 筷子, 炒锅, 扇子, 菜刀, 锅铲
家具	电视, 桌子, 椅子, 冰箱, 灶台
运动	乒乓球, 篮球, 游泳, 足球, 跑步
庆典	舞狮, 龙舟, 国旗, 月饼, 春联, 花灯
教育	铅笔, 黑板, 毛笔, 粉笔, 原子笔, 剪刀
乐器	古筝, 二胡, 唢呐, 鼓, 琵琶
艺术	书法, 皮影, 剪纸, 秦始皇兵马俑, 鼎, 陶瓷

Table 1: 图片类别列表

排序。图文检索任务的数据样例如图1a所示。该任务主要评价的是模型的图文表示对齐能力。该任务在标注时，我们要求每个标注者对给定的图片写5句不同的文本进行描述，同时要求不同文本之间的重复应少于30%。

2.3 视觉问答

视觉问答任务定义为给定一张图片及一个与图片内容相关的问题，要求模型根据图片使用短语给出答案。视觉问答任务的数据样例如图1b所示。该任务主要评价的是模型的图文理解和简单推理能力。在数据集中，每张图片对应3个问题和答案对。该任务在标注时，我们要求每个标注者对给定的图片写3个问题，并用尽可能简洁的短语给出正确答案。

2.4 视觉定位

视觉定位任务定义为给定一张图片及一个描述图中实体的短语，要求模型在图中画出该实体对应的边界框（bounding box）。视觉定位任务的数据样例如图1c所示。该任务主要评价的是模型的图片细节理解和区分能力。为了使该任务更具挑战性，该任务数据集中每张图片都对应若干句描述同一类实体不同个体的文本。比如例子中的两个文本就分别描述了图中各不相同的两个皮影。这样的任务设置方式，显然比让模型分辨图片中的皮影和人这种不同类别的实体更具挑战性。该任务标注分为两个阶段，第一阶段要求标注者将图片中我们定义的92类实体都用边界框框出来。第二阶段，我们将标注者分为两组，其中第一组标注者为图片中与图片类别实体相同的每个实体写一句自然语言描述文本，用以将其和其他实体区分开。（例如，对于一张在狗类别中的图片，第一组标注者要为图中的每条狗都写一句描述，用以区分当前描述的狗和其他的狗。）而另一组标注者则只能看到图片和描述文本，这些标注者要根据文本在图中点出文本描述的实体。如果实体在正确的边界框内，则视为该标注正确，否则视为标注错误，要求其他标注者重新标注。

2.5 视觉对话

视觉对话任务定义为给定一张图片、一段对话历史和一个问题，要求模型根据对话历史和当前的问题从100个候选答案中选出可能性最高的10个答案并对它们进行排序。（这100个候选答案是参考Das等人 (2017)的工作，从所有答案中选出的）视觉对话任务的数据样例如图 1d所示。该任务主要评价的是模型的图片整体理解、对话历史理解、指代消解和文本生成的综合能力。视觉对话任务与视觉问答任务的区别主要有两点：一是视觉问答要求模型用尽可能简单的短语回答问题，而视觉对话则要求模型回复完整的句子；二是视觉问答一般是较为直观的问题，不涉及历史信息，而视觉对话任务还要求模型对历史对话信息有所理解，里面还涉及到指

1.桌子中间摆放着火锅

2.两种口味的火锅摆放在木质的桌子上

3.一个辣的和一个菌汤锅底的火锅放在桌上

4.火锅四周摆满了涮火锅用的蔬菜、肉、丸子等食材

5.桌子中间摆放着两个口味的火锅，周围的陶瓷碗里盛放着涮火锅用的食材

(a) 图文检索

Q: 龙舟划向什么方位?

A: 右方

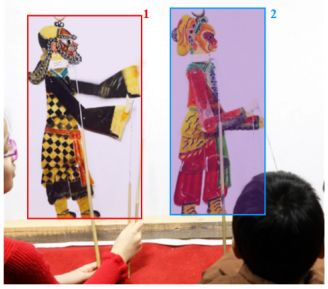
Q: 有几支队伍在划龙舟?

A: 5

Q: 大多数人的姿势是站立还是坐着?

A: 坐着

(b) 视觉问答



1.戴眼镜女孩手里拿着的皮影

2.短发男孩手里拿着的皮影

(c) 视觉定位

Caption: 蓝色桌垫上有许多食物

Q1: 桌上都有哪些食物?

A1: 食物中有鸡蛋、包子、小菜、馒头和粥

Q2: 桌上的粥是哪种粥?

A2: 桌上的粥是黑米粥

.....

Q10:桌面上的鸡蛋有几个?

A10:桌面上有两个鸡蛋

(d) 视觉对话

Figure 1: 评测中各子任务数据样例.

代消解的能力。显然，视觉对话是对模型各项能力的一个综合评价，相比视觉问答对于模型能力有更高的要求。该任务在标注时，我们将标注者分为两组，其中第一组能看到当前图片，而第二组只能看到一句描述当前图片的文本（来自图文检索标注）。第二组标注者要对第一组标注者就图片内容进行提问，以此尽可能想象出当前的图片，而第一组标注者需要根据图片据实回答。每张图片要求进行10轮问答。

2.6 数据统计

任务	训练集	开发集	测试集
图文检索	17,920	3,116	8,973
视觉问答	43,086	7,713	21,507
视觉定位	28,950	5,196	14,497
视觉对话	39,750	6,510	20,360

Table 2: 各子任务数据集样本数量统计

表2中列出了本次任务中各子任务数据集中样本数量的统计。其中图文检索任务中列出的是图片的数量，每张图片对应了5个文本描述。视觉对话列出的是对话轮次数量，该任务中每张图片对应10轮对话，我们将每轮对话及其历史作为一个样本。

3 评价指标

本次评测中，针对不同子任务，我们选择了不同的评价指标。具体来说，对于图片检索、文本检索及视觉对话，由于都是从候选集合中选出目标并进行排序，因此我们使用前1/前5/前10的召回率（$R@1/R@5/R@10$）作为评价指标，计算方式如下：

$$R@N = \frac{\text{正确结果在前}N\text{个出现的样本数}}{\text{总样本数}}$$

对于视觉问答任务，我们使用预测结果的正确率（Accuracy）作为评价指标，计算方式如下：

$$Accuracy = \frac{\text{预测正确的样本数}}{\text{总样本数}}$$

对于视觉定位任务，我们使用预测边界框和正确边界框的重叠度（Intersection over Union，IoU）作为评价指标，计算方式如下：

$$IoU = \frac{\text{正确边界框和预测边界框重叠部分面积}}{\text{正确边界框面积} + \text{预测边界框面积} - \text{二者重叠部分面积}}$$

最终，我们计算上述所有指标的宏平均值作为最终排名的依据。

4 提交结果

参赛队伍	TR			IR			VQA	VG	VD			AVG
	R@1	R@5	R@10	R@1	R@5	R@10	Acc	IoU	R@1	R@5	R@10	
江南大学	40.7	61.3	66.0	28.5	46.4	50.7	39.0	48.9	**30.1**	**44.8**	**51.9**	46.2
哈工大（深）	**43.8**	**69.4**	**78.4**	28.7	52.9	63.8	47.9	10.8	23.9	37.5	44.1	45.6
北语	41.9	68.9	78.1	**28.9**	**54.1**	**65.1**	**55.6**	**49.3**	29.5	43.8	50.7	**51.4**

Table 3: 本次评测所有提交队伍结果，其中TR表示文本检索任务，IR表示图片检索任务，VQA表示视觉问答任务，VG表示视觉定位任务，VD表示视觉对话任务，AVG表示各指标平均值，Acc表示正确率，IoU表示重叠度。参赛队伍中哈工大（深）表示哈尔滨工业大学（深圳），北语表示北京语言大学。

本次评测任务共有11支队伍报名参赛，由于任务较为困难，最终只有3支队伍提交了结果，分别是江南大学、哈尔滨工业大学（深圳）和北京语言大学。本次任务评测阶段采取盲测方式，即在测试阶段我们将没有答案的5个子任务的测试集问题及对应图片发给各支队伍，各支队伍使用各自训练好的模型预测出答案之后统一提交给评测组织方。然后由我们统一对各队伍提交的结果计算各子任务得分，并将所有指标的平均值作为最终排名的依据。上述三支队伍提交的结果见表3，各指标上最高的结果用加粗字体标出。可以看出，三支队伍各自在部分子任务上取得了最好成绩，其中，江南大学队伍在视觉对话子任务上取得最好成绩，哈尔滨工业大学（深圳）队伍在文本检索子任务上取得最好成绩，而北京语言大学队伍在图片检索、视觉问答和视觉定位三项子任务上都取得了最好成绩。最终，北京语言大学队伍凭借51.4的平均分数取得了本次评测的综合最好成绩。

同时，值得注意的是，三支参赛队伍在五个子任务上得到的结果的绝对值都比较低，尤其是其中的图片检索、视觉问答、视觉定位和视觉对话几个任务。这种结果一方面证实了该评测任务具有较大挑战性，另一方面也说明目前的图文预训练模型在面对中文图文理解问题时，仍然有巨大的进步空间。

5 方法概述

本节中我们将分别介绍三支队伍所使用的方法。

5.1 江南大学队伍

江南大学队伍采取了直接使用任务提供的训练集在X^2VLM (Zeng et al., 2022)预训练模型上针对每个子任务单独进行微调的方法。X^2VLM模型采用模块化架构，包括视觉编码器、文本编码器和多模态融合模块，所有模块基于Transformer架构。其特点是能够在预训练过程中同时学习多粒度的视觉-语言对齐和定位，支持图像-文本和视频-文本任务的统一预训练，并且具有高适应性，可以通过替换文本编码器适应不同语言或领域的任务。与CLIP (Radford et al., 2021)主要学习全局图像和文本特征不同，X^2VLM还学习对象和区域级别的细粒度特征对齐，展示了更强的多任务处理能力和灵活性。组织方提供的X^2VLM模型[1]是在中文的图文对数据上进行预训练得到的。

5.2 哈尔滨工业大学（深圳）队伍

而哈尔滨工业大学（深圳）队伍也采用了类似的微调策略，但他们同时还对比了多个支持中文的图文预训练模型上微调的结果，具体包括：

[1]https://github.com/zengyan-97/X2-VLM

- CCLM (Zeng et al., 2023)：该模型是一个跨视图语言建模框架，旨在统一跨语言和跨模态的预训练，其特点在于同时处理多模态数据（如图像-字幕对）和多语言数据（如平行句对），通过条件掩码语言建模、对比学习和匹配目标，最大化不同视图之间的互信息，从而将它们对齐到一个共同的语义空间。

- 中文CLIP (Yang et al., 2022)：该模型是一个专门针对中文的视觉-语言基础模型，通过两阶段预训练方法实现。在第一阶段中，模型采用锁定图像编码器，仅优化文本编码器的方式进行训练；在第二阶段，解锁图像编码器，进行对比学习，从而使整个模型适应中文数据集。

- OFA (Wang et al., 2022)：该模型是一种任务无关和模态无关的框架，旨在通过一个简单的序列到序列学习框架统一架构、任务和模态。OFA在预训练和微调阶段采用基于指令的学习，不需要为下游任务添加额外的任务特定层。OFA模型能够有效地转移到未见过的任务和领域，具有出色的零样本学习能力和域适应能力。

其中，中文CLIP模型只在图文检索任务中使用，而OFA模型只在图文检索和视觉定位任务中使用。根据开发集上的结果，他们最终选取X^2VLM模型用于图文检索、视觉定位任务，CCLM模型用于视觉问答、视觉对话任务。

5.3 北京语言大学队伍

北京语言大学队伍采用了拉近真实图片与答案导向图片的中文图文多模态理解增强方法。他们首先使用VisCPM-Paint模型 (Hu et al., 2023)的文生图功能，利用文本标注信息生成答案导向的图片，从而对训练数据进行扩充。然后，生成的图片与原始图片一起送入模型进行微调，对不同子任务设置不同的训练目标，拉进生成的图片与真实图片距离。

具体来说，在第一步文生图阶段，该队伍针对不同子任务使用不同的提示词进行生成。例如，对于图文检索任务，将每张图片对应的五个描述文本拼接起来作为提示词来生成和原始图片类似的图片。对于视觉问答任务，将每张图片对应的3个问题和答案组合，来生成符合所有问题、答案的图片。对于视觉对话任务，将每张图片对应的10轮问答中的答案进行组合，来生成符合对话的图片。

在第二步精调阶段，该队伍针对不同任务采用了不同的精调模型架构。针对图文检索任务，采用的精调模型通过共享参数的视觉编码器编码真实图片和答案导向图片，并使用余弦向量损失和KL散度损失分别拉近它们的表示和相似度矩阵。最后，将这些损失与原本的匹配损失和对比学习损失共同作为优化目标进行多模态微调。对于视觉问答和对话任务，该队伍计算真实图片和答案导向图片结果的下一个符号预测损失，并通过KL散度损失拉近它们输出解码器的概率分布。而对于视觉定位任务，则通过损失函数和KL损失缩小答案导向图片和原图的预测边界框与目标边界框之间的差异。

6 总结

本次任务针对目前中文上缺少全面的，不包含西方文化偏置图片的图文多模态评测数据集的问题，开展了包含图片检索、文本检索、视觉问答、视觉定位和视觉对话等五个子任务的中文图文多模态理解评测。本次评测吸引了11支来自学术界和工业界的队伍报名，但由于任务难度较大，最终只有3个队伍提交了结果。提交结果的队伍使用本任务提供的数据对多个现有的中文图文预训练模型进行了评测，同时还对基于文生图的数据增广方法进行了探索。从方法上来看，提交了结果的队伍主要方法集中在使用主办方提供的数据进行精调上。从最终结果来看，所有队伍的结果的绝对值都比较低，这一方面反映了本评测任务的挑战性，另一方面也说明目前中文图文预训练模型在实际的、没有西方文化偏置的图片上的性能仍然比较弱，还有较大进步空间。该结果也说明，解决文化偏置问题是未来中文图文预训练模型的一个重要发展方向。本次评测虽然包括了5个子任务，但在图片使用上实际不同任务之间是有比较多重叠部分的，即很多图片都有不止一种任务的标注。这种特性可以较好地支持多任务学习，但遗憾的是，在参赛队伍提交的方案中，我们没有发现对于不同任务之间关系及图文预训练模型中多任务学习方法的探索。我们认为在未来，这种图文多任务学习是一个值得探索的方向。

第二十三届中国计算语言学大会论文集，第364页-第371页，太原，中国，2024年7月25日至28日。
卷3：评测报告

参考文献

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1080–1089. IEEE Computer Society.

Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 52–59. Computer Vision Foundation / IEEE.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *CoRR*, abs/2308.12038.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Trans. Multim.*, 21(9):2347–2360.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 171–181. The Association for Computational Linguistics.

Pierre Stock and Moustapha Cissé. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 504–519. Springer.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 217–223. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6418–6428. Association for Computational Linguistics.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: contrastive vision-language pretraining in chinese. *CoRR*, abs/2211.01335.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale japanese image caption dataset. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 417–421. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022. X^2-vlm: All-in-one pre-trained model for vision-language tasks. *CoRR*, abs/2211.12402.

Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng, and Xinsong Zhang. 2023. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5731–5746. Association for Computational Linguistics.