ACL'21 | 对话系统也要进军多模态了！

原创 胡煌，梁祖杰 夕小瑶的卖萌屋 2021-08-30 22:20

微信扫一扫
关注该公众号

文 | 胡煌，梁祖杰
编 | 小轶

对于一个对话Bot来讲，拥有对视觉信息的感知和联想能力是非常重要的。比如，我们人类在对话中谈到大海的时候，就会自然地联想到蓝天、白云和飞翔的海鸟。然而，当前的对话模型，如 Meena、BlenderBot、DialoGPT 等，都是在纯文本语料上进行训练得到的，在学习过程中，缺乏对视觉信息的感知和理解。因此，基于图像的对话任务（Image-Grounded Conversation）被提出来关注这个挑战。现有的工作侧重于，探索基于给定图像的多模态对话模型。也就是说，这些工作都假设整个对话是围绕一张给定的图片进行展开的。然而，人类之间的对话是在某个特定的时刻，根据聊天的内容联想到物理世界中相关的视觉信息的。因此，这篇论文研究了开放式的基于图像的对话，即假设没有成对的对话和图像数据。具体来说，作者们提出了一种神经对话模型 Maria，可以**从大规模图像数据中检索出符合对话语境的视觉信息，来进行对话的回复**。大量实验表明，Maria 在自动和人工评估中显著优于现有的 SOTA 模型，并且可以生成一些具有视觉常识的对话回复。

论文标题：
Maria: A Visual Experience Powered Conversational Agent

论文链接：
https://arxiv.org/abs/2105.13073

Github 链接：
https://github.com/jokieleung/Maria

arxiv访问慢的小伙伴也可以在 **【夕小瑶的卖萌屋】**订阅号后台回复关键词 **【0830】** 下载论文PDF~

背景介绍

尽管最近在纯文本语料库上训练的大规模对话模型，如 Meena、Blender 和 DialoGPT 已经展现出了惊艳的表现。但它们在学习过程中，仍然缺乏对现实世界的视觉感知能力。最近，Bisk等人发表在 EMNLP 2020 上的一项研究工作 [1] 指出，共同的世界体验是让语言交流真正有意义的关键所在。另一项先前的研究 [2] 也表明，视觉感知是一种丰富的信号，可以用于建模仅仅通过文本无法记录的世界体验。图1展示了一个对话的例子，当人类 B 在谈论夏威夷海滩的假期时，人类 A 联想到了过去自己在沙滩上打排球或烧烤的经历。然而，沙滩和排球（或烧烤）之间的关联关系在传统知识库中（例如，知识图谱）却很难捕捉到。


Human-A: Hey! How was your vacation?

Human-B: *Awesome! I had a good time with my friends in Hawaii, the beaches are very beautiful there.*

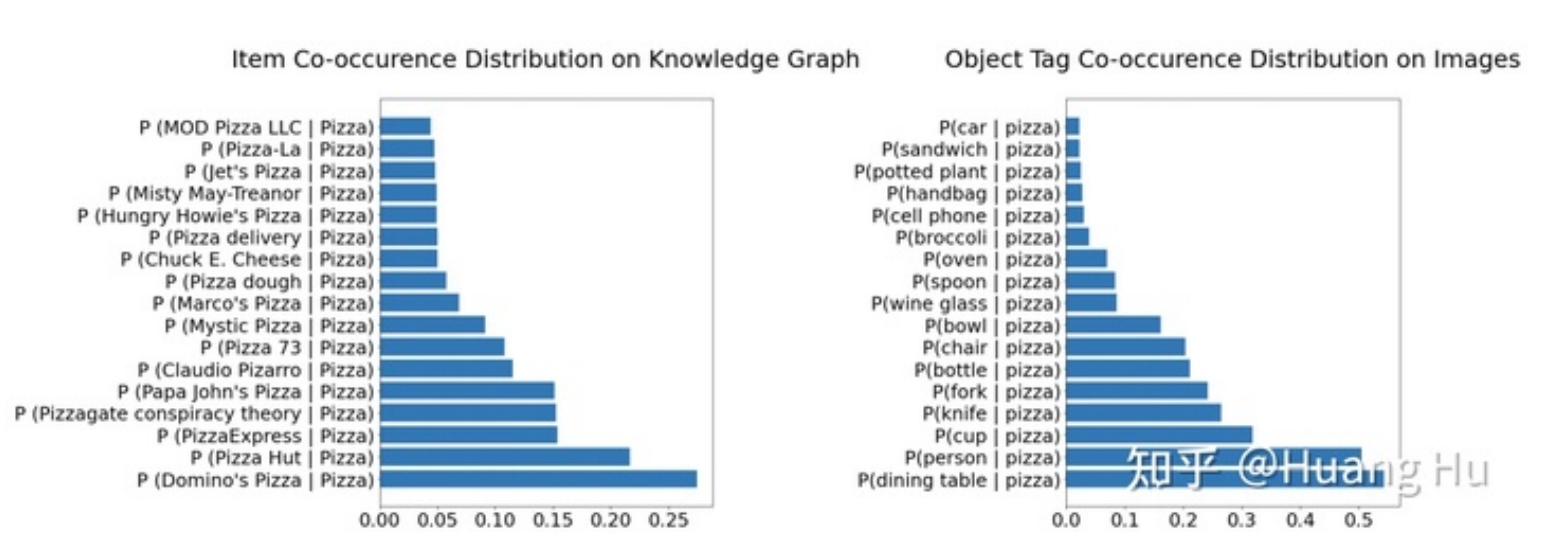
Human-A: Cool! did you play beach volleyball with your friends?

(**Human-A:** Cool, have you had a BBQ with your friends on the beach? The grilled fish was great!)

Human-B: *Nope, but it sounds great. Maybe next time.*

▲人和人之间对话的例子。当B谈论在夏威夷海滩上的假期时，A想起了自己曾经在海滩上打沙滩排球（或者进行海滩烧烤）的经历

受此启发，本文选取了一个常见单词“pizza”，并收集了在 Google 知识图谱 [3] 和 MS-COCO [4] 图像数据上与“pizza”共现次数最多的 17 个词。如图2所示，知识图谱上与“pizza”共现的词往往是一些抽象概念，比如“Pizza Hut”（必胜客），而图像数据中物体标签的共现关系反映了我们现实世界的一些常识，例如“pizza”通常在“餐桌”上，人们在吃“pizza”时通常使用“刀”。有趣的是，我们还发现“pizza”也与“手机”甚至“植物”也经常共同出现。这说明人们在吃“pizza”时，有时会把“手机”放在桌子上，或者餐厅里可能摆放了一些“植物”。更多例子参见论文附录。



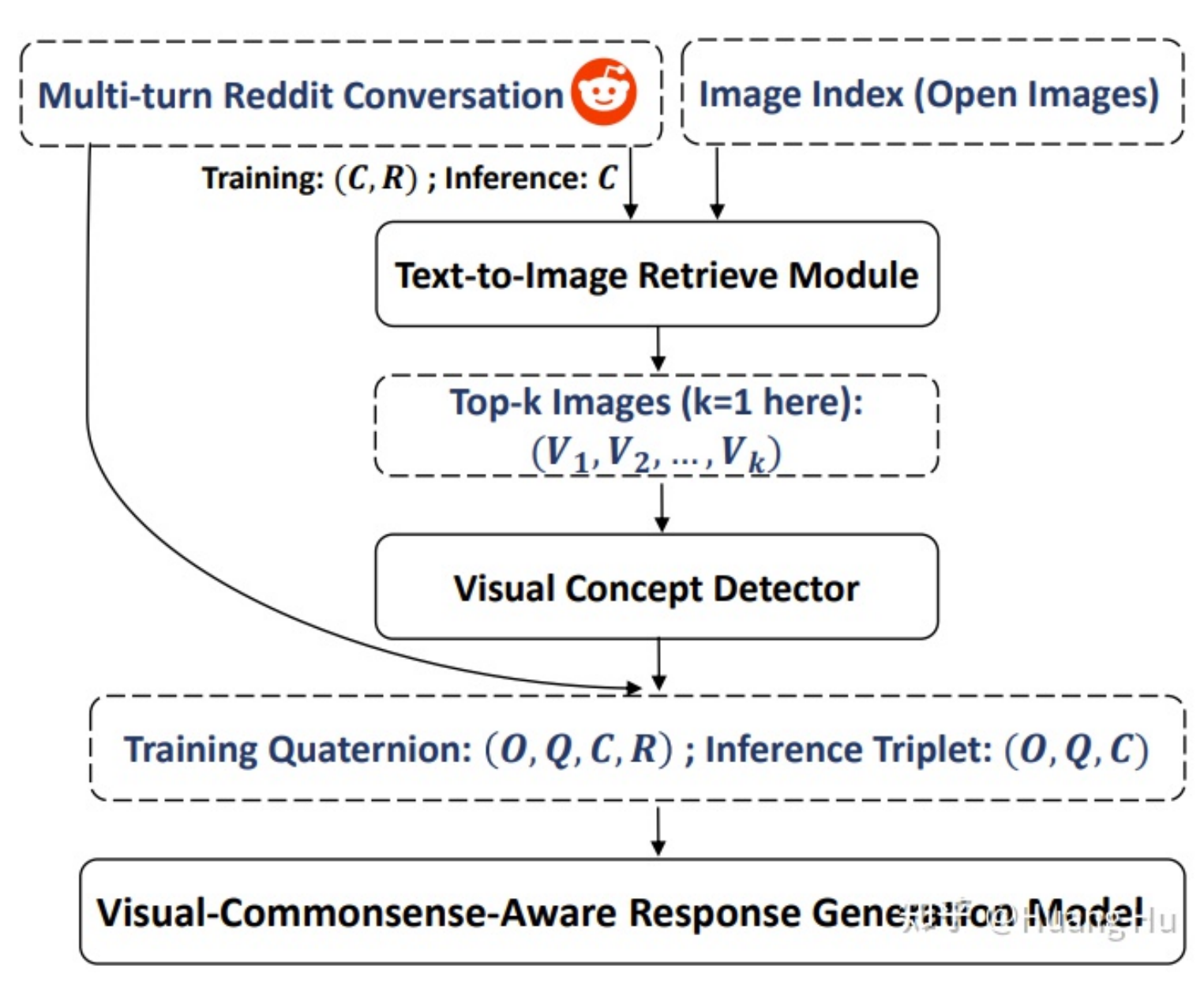
▲图2：在Google知识图谱和MS-COCO图像数据集上，与Pizza共现的词的分布

因此，赋予对话模型对物理世界的视觉感知能力，对帮助它们真正地理解对话语境至关重要。

这篇论文研究在没有给定图像的情况下，如何利用非平行的视觉信息来辅助对话的场景，即没有成对的对话和图像数据。具体地，本文提出了 Maria，一种由视觉体验驱动的神经对话模型，这些视觉信息是从预先构建的图像索引库（如，Open Images [5]）中检索出来的。

具体方法

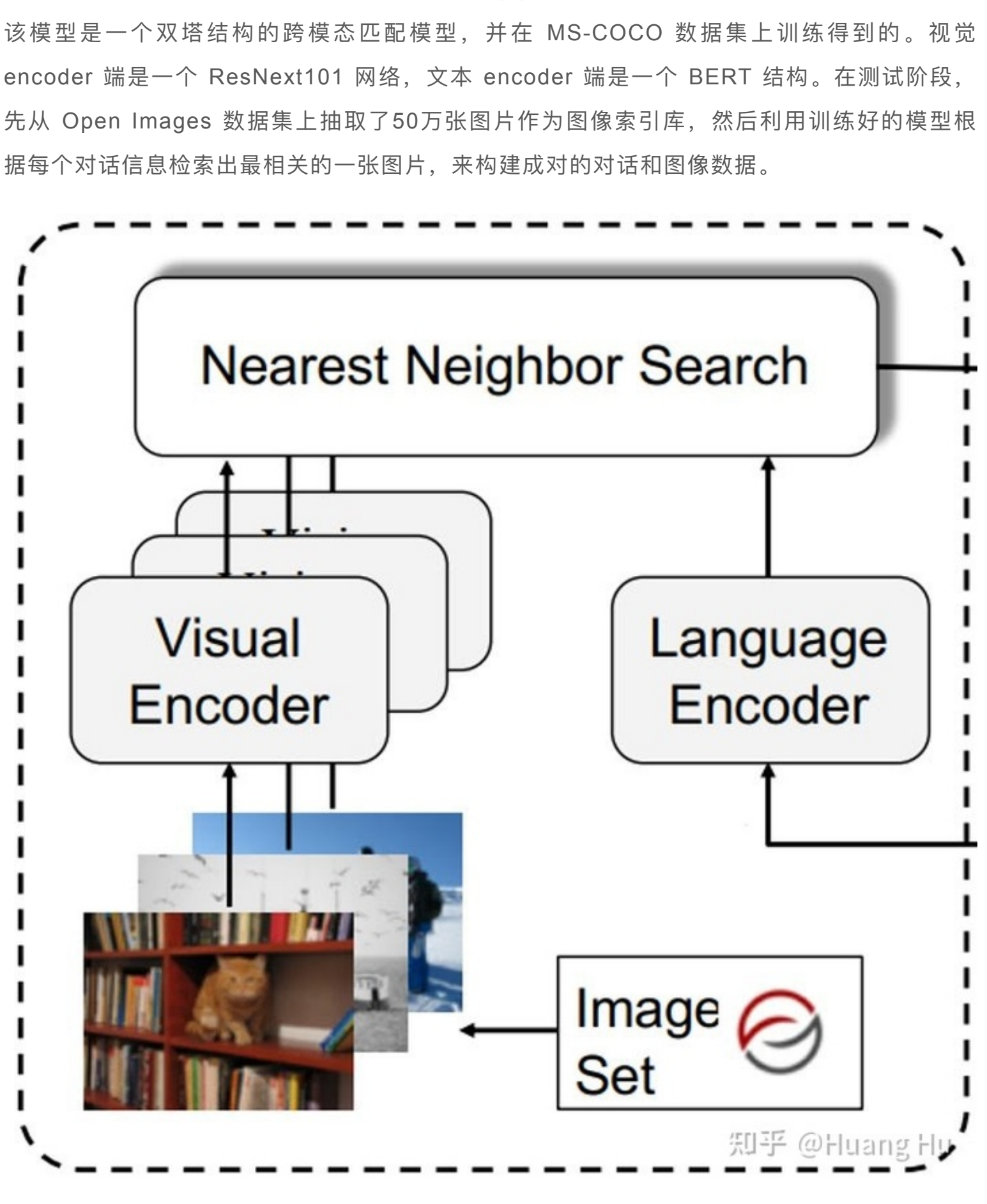
Maria 由三个组件组成：图像检索模型、视觉概念检测模型和基于视觉知识的对话生成模型。图像检索模型负责根据对话的上下文检索出可视化世界体验，例如，从图像数据库中检索出与对话相关的图像。图3是Maria框架的流程图。



▲图3：Maria框架的流程图

1. 文本到图像的检索器 (Text-to-Image Retriever)

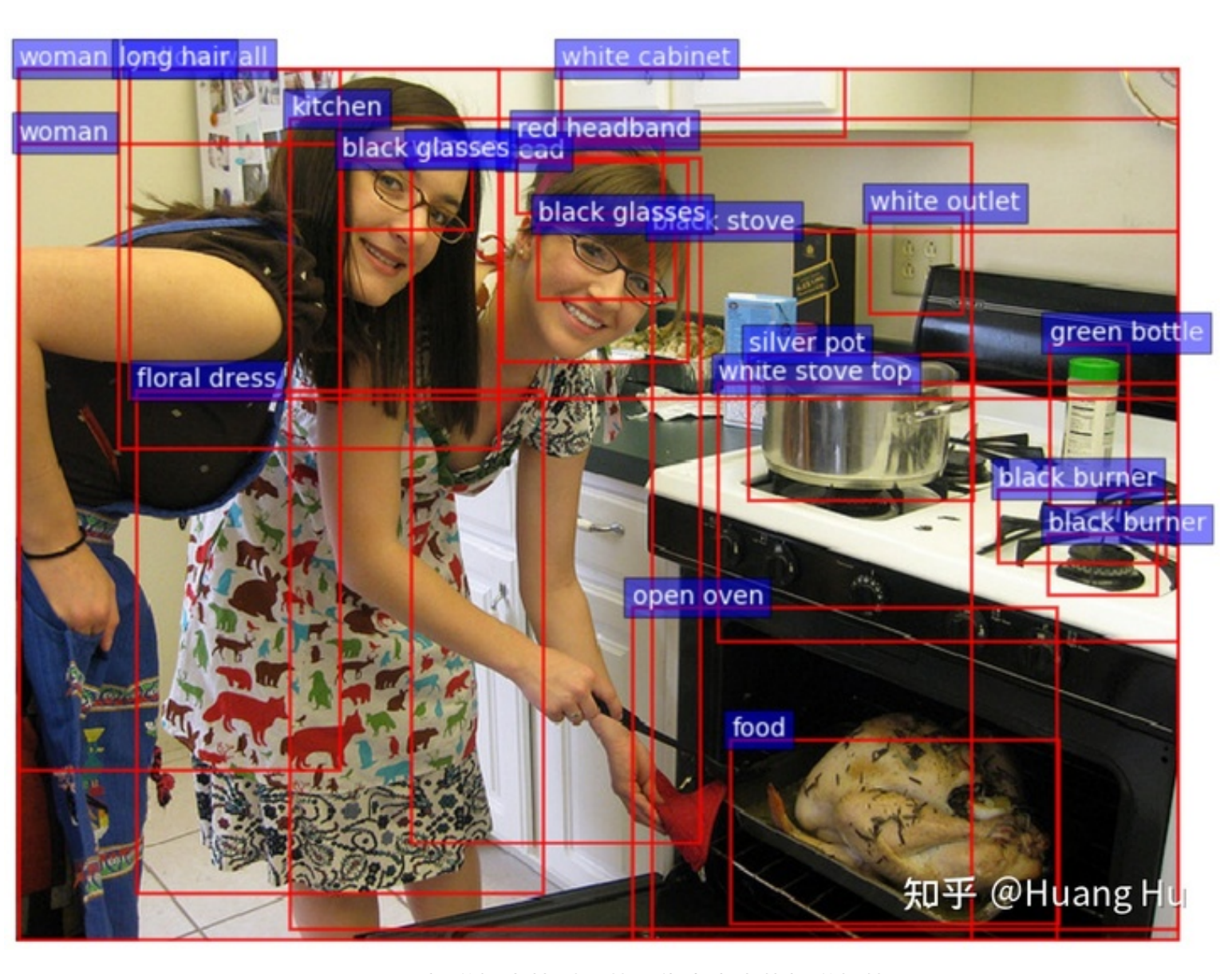
对于图像检索模型，本文复用了 Vokenization [6] 中的文本到图像的检索模型。如图4所示，该模型是一个双塔结构的跨模态匹配模型，并在 MS-COCO 数据集上训练得到的。视觉 encoder 端是一个 ResNext101 网络，文本 encoder 端是一个 BERT 结构。在测试阶段，先从 Open Images 数据集上抽取了50万张图片作为图像索引库，然后利用训练好的模型根据每个对话信息检索出最相关的一张图片，来构建成对的对话和图像数据。



▲图4：文本到图像的检索模型架构图 (from Vokenization)

2. 视觉概念检测器 (Visual Concept Detector)

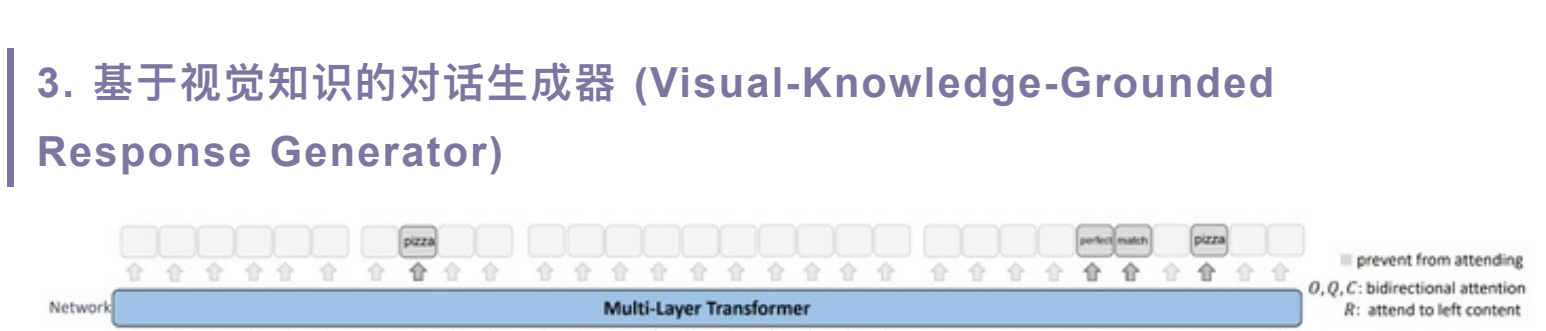
给Reddit语料中的每个对话匹配了一个最相关的图片之后，视觉概念检测模型，即为UpDown [7] 在 Visual Genome [8] 数据集上预训练的物体检测器，从检索出来的图片中提取图像的区域特征（bboxes）和相应的视觉概念标签（object tags），比如图5中显示的蓝色标签：kitchen, black glasses, open oven 等。



▲图5：视觉概念检测器从图像中出去的视觉标签

最终，我们可以构建（bboxes、tags、context、response）四元组作为后续多模态对话生成模型的训练数据。

3. 基于视觉知识的对话生成器 (Visual-Knowledge-Grounded Response Generator)



▲图6：对话生成模型的架构图

图6是对话回复生成模型的结构，在输入层面引入了四种 embedding，即为 Token-Level、Turn-Level，Position-Level、Segment-Level。为了让对话模型有效地学习到从图像数据中抽取的视觉知识，本文提出了以下三个任务：

1) Masked Concept Prediction (MCP)

由于视觉和文本两种模态之间的语义偏差，直接通过回复生成目标来优化模型会使得对视觉信息的利用不够充分。为了对齐两种模态的语义表示，作者提出了Masked Concept Prediction目标：在训练过程中，随机 mask 15%的 visual concept 让模型来预测。由于视觉概念标签从图像中抽取时，没有明确的顺序，所以需要把该任务建模成一个预测集合的匹配问题，并且利用 Hungarian Matching Loss [9][10] 作为优化目标。

输入定义为 $X=(O,Q,C,R)]$，其中 O 表示提取的 region 序列，Q 表示对应的标签序列，C 表示对话上下文，R 表示回复。其中，双向自注意力的部分为 $B=(O,Q,C)$，masked 视觉概念集合为 $\hat{Q}$，没有mask的部分表示为 $B\setminus\hat{Q}$，对应的预测概率表示为 $H=\{h_i\}_{i=1}^{m}$。其中 h_i 是第 i 个 mask 位置预测的概率分布。那么，MCP 任务的损失函数可以定义为：

$$\mathcal{L}_{\mathrm{MCP}}(\hat{Q},H,\alpha)=-\sum_{q_{\alpha(i)}\in\hat{Q}}\log h_i\left(q_{\alpha(i)}\mid B\setminus\hat{Q}\right)$$

其中，$\alpha(i)$ 是第 i 个预测位置上目标视觉概念标签的索引。因此，当模型在预测 mask 的视觉概念时，会学习其与 region 特征、对话上下文以及其他没有 mask 的视觉概念之间的对应关系。

2) Masked Response Prediction (MRP)

为了让模型更好地利用不同模态的输入信息，作者采用了 UniLM [11] 中的 Mask Response Prediction（MRP）来建模对话生成任务，即在训练的时候是Masked Language Model (MLM) 任务。回复中被mask的tokens表示为 $\hat{R}$，其它所有的输入表示为 $X\setminus\hat{R}$。假设 p_i 是 R 中第 i 个 token，MRP 的损失函数可以定义为：

$$\mathcal{L}_{\mathrm{MRP}}(X,\hat{R})=-\sum_{w_i\in\hat{R}}\log p_i\left(w_i\mid X\setminus\hat{R}\right)$$

在 R 中的 self-attention mask 是自左向右的，其它的segment是双向的。也就是说 R 中的 tokens 可以 attend O,Q,C 中所有的token，以及 R 中左边的 token，这么设计是为了鼓励模型在生成回复的时候，去学习所有输入 token（包括视觉模态和文本模态）之间的对应关系。

3) Visual Knowledge Bias (VKB)

最后，引入视觉知识的偏置让模型在生成过程中，更多地生成与视觉信息相关的 token。具体地，视觉词表的偏置 b_q 定义为：

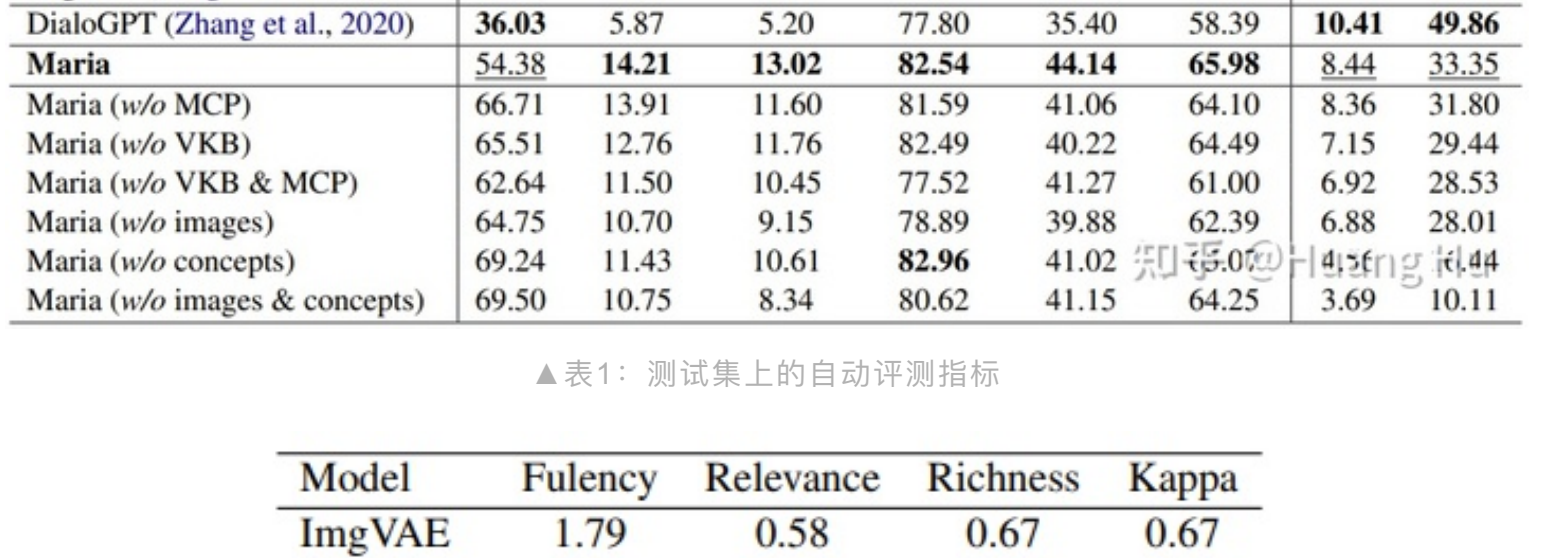
$$b_q=F_q(e^q_{avg})$$

其中，$F_q:\mathbb{R}^d\rightarrow\mathbb{R}^{|V|}$ 是一个投影层，e^q_{avg} 表示所有视觉概念隐层表示的 average pooling。之后，b_q 被加到生成模型的 softmax layer 来得到最终在词表上的概率分布：

$$\hat{\boldsymbol{p}}=softmax(We^r+b+\hat{b}_q)$$

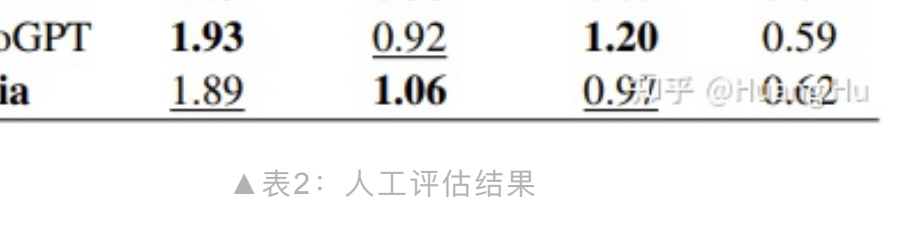
实验结果

为了评估Maria的表现，我们在Reddit Conversation Corpus [12] 上进行了实验，训练集/验证集/测试集的大小分别是1百万/2万/2万个对话，每个对话大概有3至5轮。此外，从 Open Images dataset 中抽取了50万张图片作为图像索引库，然后利用在 MS-COCO 上训练好的检索模型，给每个对话检索出一张最相关的图片，来构造成对的 dialog-image 数据。



Model	PPL	BLEU-1	Rouge-L	Average	Extrema	Greedy	Dist-1	Dist-2
Seq2Seq (Bahdanau et al., 2015)	77.27	12.21	10.81	78.38	40.06	62.64	0.53	1.96
HRED (Serban et al., 2016)	84.02	11.68	11.29	75.54	37.49	60.41	0.89	3.21
VHRED (Serban et al., 2017)	78.01	12.22	11.82	75.57	39.24	62.07	0.87	3.49
ReCoSa (Zhang et al., 2019)	71.75	12.75	11.75	79.84	42.29	63.02	0.66	3.83
ImgVAE (Yang et al., 2020)	72.06	12.58	12.05	79.95	42.38	63.55	1.52	6.34
DialoGPT (Zhang et al., 2020)	**36.03**	5.87	5.20	77.80	35.40	58.39	**10.41**	**49.86**
Maria	54.38	**14.21**	**13.02**	**82.54**	**44.14**	**65.98**	8.44	33.35
Maria (w/o MCP)	66.71	13.91	11.60	81.59	41.06	64.10	8.36	31.80
Maria (w/o VKB)	65.51	12.76	11.76	82.49	40.22	64.49	7.15	29.44
Maria (w/o VKB & MCP)	62.64	11.50	10.45	77.52	41.27	61.00	6.92	28.53
Maria (w/o images)	64.75	10.70	9.15	78.89	39.88	62.39	6.88	28.01
Maria (w/o concepts)	69.24	11.43	10.61	**82.96**	41.02	65.07	4.56	16.44
Maria (w/o images & concepts)	69.50	10.75	8.34	80.62	41.15	64.25	3.69	10.11

▲表1：测试集上的自动评测指标

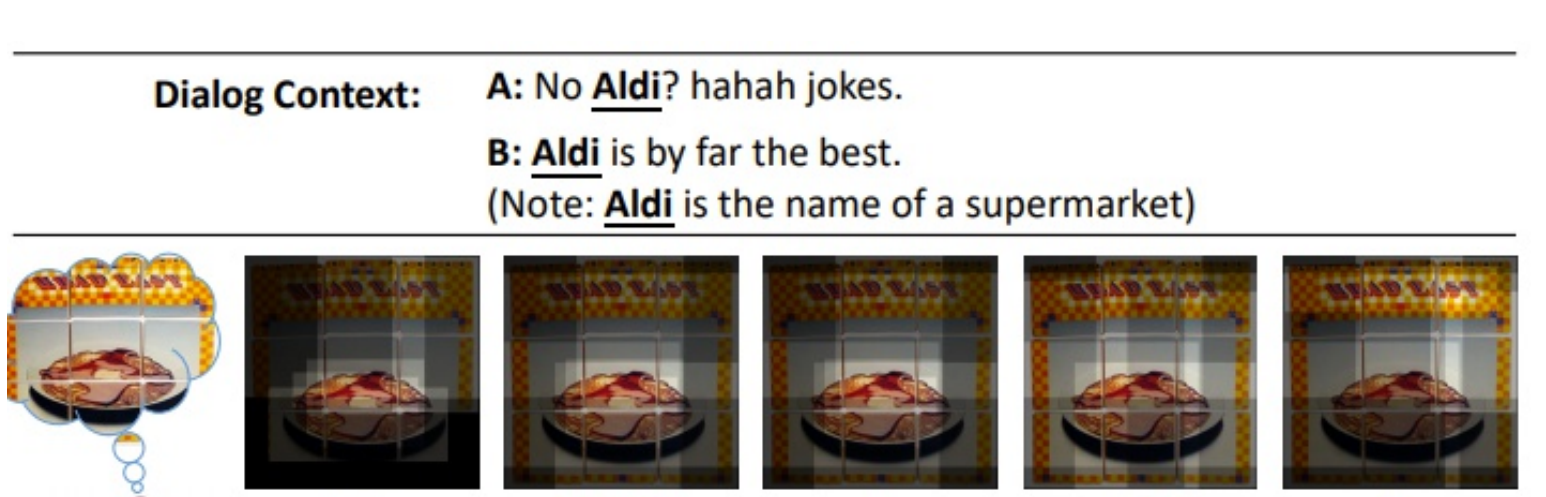


Model	Fulency	Relevance	Richness	Kappa
ImgVAE	1.79	0.58	0.67	0.67
DialoGPT	**1.93**	0.92	**1.20**	0.59
Maria	1.89	**1.06**	0.97	0.62

▲表2：人工评估结果

在表1的主实验中，本文比较了 Seq2Seq、HRED、VHRED、ReCoSa、ImgVAE，以及 DialoGPT 这些基线模型。可以看到，除了 DialoGPT 外，Maria 生成的回复质量在流程度（PPL）、相关性（BLEU/Rouge-L/Average/Extrema/Greedy）以及多样性（Dist-1/2）上都要显著优于基线模型。尤其是，与最新的 SOTA 模型 ImgVAE 的比较中，Maria 的 Dist-1/2 要显著优于 ImgVAE。这表明引入从图像中抽取的视觉知识，有助于生成多样性更好、信息度更丰富的回复。这与表2中的人工评估结果也是一致的，Maria 的内容丰富度要比 ImgVAE 明显好。另一方面，ImgVAE 中文本到图像的生成模型是在 ImageChat [13] 上训练，在Reddit对话数据上进行测试的。训练集和测试集数据分布的差异，导致 ImgVAE 实际中的效果并不理想。另外一个观测结果是，Maria 在 PPL 和 Dist-1/2 指标上的表现要略逊于 DialoGPT。原因是，DialoGPT 是一个大规模预训练的对话模型，并且引入了额外的反向模型来提高生成回复的多样性。

表1中的消融实验结果也验证了 MCP、VKB 的有效性。图7是一个可视化的例子，可以看出，当上下文谈到“Aldi”（一家连锁超市）的时候，Maria 能够“联想”到“Pizza”。图8/9 中是更多Maria对话的例子，具体的分析参见论文附录。


Dialog Context: **A:** No **Aldi**? hahah jokes.
B: **Aldi** is by far the best.
(**Note:** **Aldi** is the name of a supermarket)

▲图7：Maria对话的可视化例子


Dialog Context: **A:**the pass of the **world cup** (james rodriguez)
B: absoleutely gorgeous.

Baseline Response: **ImgVAE:** I love the world cup.
DialoGPT: it 's not a pass of the world cup.

Detected Concepts: ball, socks, field, shirt, grass, man, shorts, ***player***, sock

Response: he's the best ***player*** in the world

▲图8：Maria 对话例子a


Dialog Context: **A:**the pass of the **world cup** (james rodriguez)
B: absoleutely gorgeous.

Baseline Response: **ImgVAE:** I love the world cup.
DialoGPT: it 's not a pass of the world cup.

Detected Concepts: ball, socks, field, shirt, grass, man, shorts, ***player***, sock

Response: he's the best ***player*** in the world

▲图9：Maria 对话例子b

作者最近在招intern一起做研究哦，欢迎感兴趣的同学联系。

后台回复关键词【入群】
加入卖萌屋NLP/IR/Rec与求职讨论群
后台回复关键词【顶会】
获取ACL、CIKM等各大顶会论文集！

参考文献

[1]. Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. EMNLP 2020.

[2]. Stevan Harnad. 1990. The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3):335–346.

[3]. https://developers.google.com/knowledge-graph

[4]. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, ´ and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. ECCV 2014.

[5]. Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv 2018.

[6]. Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding via contextualized, visually-grounded supervision. EMNLP 2020.

[7]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. CVPR 2018.

[8]. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV 2017.

[9].Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng. 2016. End-to-end people detection in crowded scenes. CVPR 2016.

[10]. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. ECCV 2020.

[11]. Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. NeurIPS 2019.

[12]. Ze Yang, Wei Wu, Huang Hu, Can Xu, and Zhoujun Li. 2020. Open domain dialogue generation with latent images. AAAI 2021.

[13].Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. ACL 2020.

喜欢此内容的人还喜欢
若被制裁，中国AI会雪崩吗？
夕小瑶的卖萌屋