OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge

承子杰
dept. AMSS（数学与系统科学研究院）
of CAS (中国科学院)
chengzijie22@mails.ucas.ac.cn
202228000243001

I. 主要内容

文章作者认为 VQA 应该能从图像和文本的联合信息中学习推理，从而进行场景理解。然而现有的大部分 VQA 不需要外部知识（knowledge-based）和逻辑推理，仅局限于计数、目标检测等简单任务。因此作者提出了 OK-VQA 数据集，数据集中提出的问题需要结合外部知识来进行回答。此外作者提出了一种 ArticleNet 的方法用于结合外部知识，并选择了一系列 VQA 模型与其结合，给出了它们在 OK-VQA 数据集上的 BenchMark 用作后续研究作为基准。

II. OK-VQA 数据集的建立

A. 数据来源

作者对 VQA 数据集中 10000 个问题进行 Age Annotation，发现其 78% 的问题可以由 10 岁以下的儿童回答，因此作者认为 VQA 数据集大部分问题不需要背景知识。因此 OK-VQA 的图像数据来源于 MS COCO 的随机图像，选取其中 80K 张图片作为训练数据集，40K 张图片作为验证数据集，场景覆盖了 COCO 数据中 365 个场景中的 350 个。

B. 数据标注

作者将 120K 张图片在 Amazon 的 MTurk 平台上进行外包，经历了两轮标注。在第一轮标注中，作者要求标注者对于每一张图片提出一个问题，要求该问题与图像相关，且需要依赖一定的外部知识。在第二轮标注中，要求 5 个标注者分别对第一轮标注提出的问题进行回答，以确保提出的问题具有一定价值。最后再进行一次人工筛选，筛出 34,921 个问题。

考虑到数据集存在偏差，作者将第二轮回答中出现 5 个答案一致的问题和 5 个答案均不一致的问题进行删除，

图 1. OK-VQA 数据集样本示例

最终留下了 14,055 个问题。其中，9009 个作为训练问题，5046 个作为训练问题。图 1展示了数据集中的一些示例样本。

C. 数据结构

作者对数据集做了知识类别划分（Knowledge Category）。他们将问题分为 10+1 类，类别涵盖了车辆和交通（Vehicles and Transportation）、商标公司和产品（Brands, Companies and Products）、材料和衣服（Objects, Materials and Clothing）、运动和娱乐（Sports and Recreation）、烹饪和食品（Cooking and Food）、历史文化与地理（Geography, History, Language and Culture）、日常生活与动植物（People and Everyday Life, Plants and Animals）、科学与技术（Science and Technology）、气候环境（Weather and Climate）10 个类别和一个其他（Other）类别，各类别在数据集中的分布情况见图 2。

此外，作者还进行了数据集问题统计 (Question Statistics)。经过作者统计，在提出的 140,55 个问题中有 12,591 种问题类别（即非重复问题数量），问题题干涉及了 7178

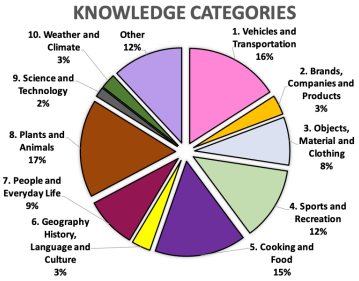

图 2. 各类别占比

Knowledge Category	Highest relative frequency question words	Highest relatively frequency answers
1. Vehicles and Transportation	bus, train, truck, buses, jet	jet, double decker, take off, coal, freight
2. Brands, Companies and Companies	measuring, founder, advertisements, poster, mobile	ebay, logitech, gift shop, flickr, sprint
3. Objects, Material and Clothing	scissors, toilets, disk, teddy, sharp	sew, wrench, quilt, teddy, bib
4. Sports and Recreation	tennis, players, player, baseball, bat	umpire, serve, catcher, ollie, pitcher
5. Cooking and Food	dish, sandwich, meal, cook, pizza	donut, fork, meal, potato, vitamin c
6. Geography, History, Language and Culture	denomination, nation, festival, century, monument	prom, spire, illinois, past, bern
7. People and Everyday Life	expressing, emotions, haircut, sunburned, punk	hello, overall, twice, get married, cross leg
8. Plants and Animals	animals, wild, cows, habitat, elephants	herbivore, zebra, herd, giraffe, ivory
9. Science and Technology	indoor, mechanical, technology, voltage, connect	surgery, earlier, 1758, thumb, alan turing
10. Weather and Climate	weather, clouds, forming, sunrise, windy	stormy, noah, chilly, murky, oasis

图 3. 各类别高频问题答案单词

个单词。除此之外，作者还统计了 10+1 个知识类别中相对频率最高的问题单词和答案单词，统计结果见图 3。通过观察，我们发现问题单词与答案单词间具有很高的关联性，具有一定的上下位关系。

III. 数据集基准（BenchMark）

A. *ArticleNet*

作者提出了名为 ArticleNet 的框架，使得可以通过 Wikipedia 的 API 进行检索外部相关文章，以此获得非结构化外部知识。其主要框架（图 4）可以分为三步：

1) 对训练数据提供的图片使用图像和场景分类器识别图像特征单词，并与其问题单词结合组成所有可能的查询（query）；
2) 使用 Wikipedia 提供的 API 对每一个查询进行检索，并保留检索结果的第一篇文章。
3) 对每个查询检索的文章，根据查询词在句子中出现的频率，选择整篇文章中最符合的句子，从而用句子来代替整篇文章的查询结果。

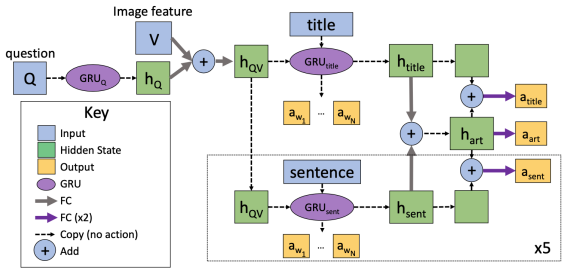

图 4. ArticleNet 基本框架

ArticleNet 也可以与其他 VQA 方法相结合，作者提供的结合方法是将查询结果句子的隐藏层表示与具体 VQA 模型中某一层的输出向量进行向量拼接。根据作者提供的各具体模型在 OK-VQA 数据上的表现（图**??**），考虑到 ArticleNet 是基于互联网数据，且查询结合外部知识方法略显简单粗糙，单独使用 ArticleNet 似乎效果并不理想。但是其与 MUTAN 和 BAN 等方法结合，却有显著的效果提升。

Method	OK-VQA	VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
Q-Only	14.93	14.64	14.19	11.78	15.94	16.92	11.91	14.02	14.28	19.76	25.74	13.51
MLP	20.67	21.33	15.81	17.76	24.69	21.81	11.91	17.15	21.33	19.29	29.92	19.81
ArticleNet (AN)	5.28	4.48	0.93	5.09	5.11	5.69	6.24	3.13	6.95	5.00	9.92	5.33
BAN [20]	25.17	23.79	17.67	22.43	30.58	27.90	**25.96**	20.33	25.60	20.95	**40.16**	22.46
MUTAN [4]	26.41	25.36	18.95	24.02	33.23	27.73	17.59	20.09	**30.44**	20.48	39.38	22.46
BAN + AN	25.61	24.45	19.88	21.59	30.79	29.12	20.57	21.54	26.42	**27.14**	38.29	22.16
MUTAN + AN	**27.84**	**25.56**	**23.95**	**26.87**	**33.44**	**29.94**	20.71	**25.05**	29.70	24.76	39.84	**23.62**
BAN/AN oracle	27.59	26.35	18.26	24.35	33.12	30.46	28.51	21.54	28.79	24.52	41.4	25.07
MUTAN/AN oracle	28.47	27.28	19.53	25.28	35.13	30.53	21.56	21.68	32.16	24.76	41.4	24.85

图 5. BenchMark on OK-VQA

此外，分析 BenchMark 的结果，我们发现这些方法在 OK-VQA 数据集上的结果均大幅低于在标准 VQA 数据集上的结果。由此可见，对于基于外部知识的 VQA 任务并没有一个有效的模型，基于此类问题的研究还有很大的空间进行探索。

B. 视觉特征简化（*Visual feature ablation*）

作者还对 MUTAN 做了一个简单的视觉特征简化实验，观测不同的视觉特征提取器在 OK-VQA 数据集上的结果变化。作者分别采用了 ResNet152、ResNet50、ResNet18 和 Q-Only 分别进行实验，结果如下：

Method	VQA score on OK-VQA
ResNet152	26.41
ResNet50	24.74
ResNet18	23.64
Q-Only	14.93

图 6. Visual feature Ablation

通过实验结果可以发现，基本还是遵循越深的视觉特征提取器可以在 OK-VQA 上获得越好的结果。

C. 尺度简化（*Scale ablation*）

最后，作者还进行了一个简单的尺度简化实验。通过对数据集中不同训练数据量进行训练，获得在测试数据集上的打分，结果如下所示

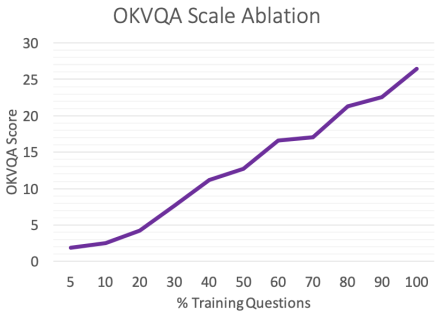

图 7. Scale Ablation

不难发现，数据量的提升可以获得更好的训练结果。通过视觉特征简化实验和尺度简化实验，我们可以初步判断 OK-VQA 是有价值的数据集。

IV. 总结

本篇文章的主要贡献主要有以下几点：

- 提出了基于外部知识的 VQA 任务；
- 构建了基于外部知识的 VQA 数据集 OK-VQA，并提供了该任务和数据集的 BenchMark
- 提出了 ArticleNet 框架，为该问题的解决提供了一种思路。
- 通过流行 VQA 模型在 OK-VQA 上的结果指明，基于外部知识的 VQA 任务没有一个有效模型。该问题具有很大的挑战和研究空间。

References

参考文献

[1] Marino K, Rastegari M, Farhadi A, et al. Ok-vqa: A visual question answering benchmark requiring external knowledge[C]//Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. 2019: 3195-3204.