The Spoon is in the Sink: Assisting Visually Impaired People in the Kitchen

Katie Baker[1], **Amit Parekh**[1], **Adrien Fabre**[1],
Angus Addlesee[1], **Ruben Kruiper**[1], **and Oliver Lemon**[1,2]
[1]Heriot-Watt University, Edinburgh, Scotland
[2]Alana AI
{kb2000,amit.parekh,abpf2000,a.addlesee,rk22,o.lemon}@hw.ac.uk

Abstract

Visual Question Answering (VQA) systems are increasingly adept at a variety of tasks, and this technology can be used to assist blind and partially sighted people. To do this, the system's responses must not only be accurate, but *usable*. It is also vital for assistive technologies to be designed with a focus on: (1) *privacy*, as the camera may capture a user's mail, medication bottles, or other sensitive information; (2) *transparency*, so that the system's behaviour can be explained and trusted by users; and (3) *controllability*, to tailor the system for a particular domain or user group. We have therefore extended a conversational VQA framework, called Aye-saac, with these objectives in mind. Specifically, we gave Aye-saac the ability to answer visual questions in the kitchen, a particularly challenging area for visually impaired people. Our system[1] can now answer questions about quantity, positioning, and system confidence in regards to 299 kitchen objects. Questions about the spatial relations between these objects are particularly helpful to visually impaired people, and our system output more *usable* answers than other state of the art end-to-end VQA systems.

1 Introduction

Visual impairment can lead to seemingly unrelated health issues. Specifically, malnutrition has been associated with visual impairment because of the difficulties encountered when shopping for, preparing, and eating food (Chung et al., 2021; Jones et al., 2019). One major issue is that preparing a meal involves various situations where visually impaired people feel unsafe. A lack of spatial awareness and depth perception, especially when using a knife or preparing hot meals, contributes to concerns about getting injured.

Another common concern is hygiene, e.g. the inability to read expiry dates, see dirt on vegetables, recognise mold on food or spot that meats are thoroughly cooked. As a result, most visually impaired people only prepare meals with the help from family members or carers — or simply do not prepare hot meals at all (Jones et al., 2019).

In this study, we explore the use of a Visual Question Answering (VQA) system to alleviate some of the issues that visually impaired people encounter in their kitchen. We describe Aye-saac, a voice assistant that can locate objects commonly found in a kitchen. The underlying architecture for Aye-saac was developed in 2020 by students at Heriot-Watt University and was designed to be both *transparent* and *controllable*, aligning with our needs. We extended the object detection capabilities from 30 to 299 types of kitchen objects. Further added functionality includes:

1. Object positioning in a scene, e.g. "*The spoon is in the sink*". We describe spatial relations to 'anchor' objects or a user's hands where possible — this avoids the use of other movable objects in the output, rendering the output useless, e.g. "*The spoon is next to the fork*".

2. Handling queries on the quantity of objects within a scene, e.g. "*I count one carrot and two fish*". This is particularly useful for counting ingredients.

3. Transparency of confidence scores for responses generated by Aye-saac, e.g. "*I am 72% certain that the spoon is in the sink*".

We investigate how well Aye-saac handles the newly added user-intents, how well it detects the newly added kitchen objects, and compare the object detection capabilities against two dedicated end-to-end (E2E) VQA systems. We show that

[1] Code and evaluation data can be found at: https://github.com/Aye-saac/aye-saac

Proceedings of the Conference on Reasoning and Interaction (ReInAct), pages 32–39
October 4-6, 2021. ©2021 Association for Computational Linguistics

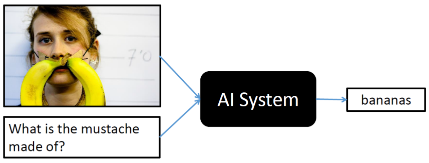


Figure 1: Example image and question input to a Visual Question Answering (VQA) system, with generated response, from Antol et al. (2015).

Aye-saac outperforms these two E2E systems when queried about object positioning by returning more descriptive and *usable* positioning information.

2 Related Work

2.1 Visual Question Answering

A VQA system takes a natural language question and an image as input, aiming to reason over the contents and respond with a natural language utterance (Antol et al., 2015). Recently, VQA systems that achieve state-of-the-art results have been trained E2E, an example is the Pythia system that won the 2018 VQA competition (Jiang et al., 2018). While these systems provide improved performance, they lack the ability to explain why they generated a specific response (Li et al., 2018) and have to be trained on large datasets.

A crowd-sourcing approach can provide VQA systems with the required training data to cover many subjects and handle common questions (Gurari et al., 2018), but the objects or properties in an image must be within this data to be mapped correctly between the query and image (Antol et al., 2015). As an example, consider Figure 1: if the model was not trained to recognise bananas, the system is unable to correctly handle queries related to bananas. This is a challenge when working within a very specific domain.

The *VizWiz Social* (Brady et al., 2013) app was used by visually impaired people to collect pictures and questions about their surroundings. The resulting VizWiz dataset (Gurari et al., 2018) contains these images and questions with answers crowd-sourced from sighted volunteers. Although the kitchen-specific subset is not large enough to train an E2E system upon, this study provides useful insight into the types of questions that are commonly asked by visually impaired people, and the kitchen was identified as a particularly challenging area. We used this dataset to direct our work.

To the best of our knowledge, there is no dataset available to train a VQA system that assists visually impaired users in a kitchen setting. Some existing datasets to train VQA systems contain subsets of data that are situated in the kitchen, such as *Embodied Questioning Answering* (EQA) (Das et al., 2018) and *Interactive Question Answering Dataset* (IQUAD) (Gordon et al., 2018). However, both EQA and IQUAD use computer-generated questions that are grounded on a synthetic 3D environment. Synthetic visual scenes do not have the randomness of real-life (Hudson and Manning, 2019) which biases the model towards certain environments. Questions also lack diversity due to a templative generation method, biasing the model further (Das et al., 2018).

The lack of a domain-specific dataset complicates the use of an E2E approach (Zhao et al., 2019). Instead, Aye-saac only relies on neural models for object detection and Natural Language Understanding (NLU). A rule-based approach is used to process the image and query, and to formulate a response in real-time. This makes it possible to tweak and extend Aye-saac in a controlled manner. Additionally, if a question is answered incorrectly, the system can provide reason for the response.

2.2 Spatial relationships

Current object detection systems perform well at detecting entities, but are not robust at inferring the spatial relationships between them (Krishna et al., 2017). This weakness stems from the available datasets, as they lack relative spatial positioning information and must implicitly infer them.

To address this limitation, Krishna et al. (2017) introduced the *Visual Genome* (VG) dataset, converting natural language descriptions of images to dense scene graphs that include spatial relationships and common descriptive attributes of entities. Datasets such as *CompGuessWhat?!* (CGW) (Suglia et al., 2020) and GQA (Hudson and Manning, 2019) extend object detection datasets by including dense scene graphs that contain additional situational and abstract attributes, and further include binary question-answer pairs grounded on the context of the scene (Suglia et al., 2020; Hudson and Manning, 2019). However, models trained on these datasets are not as robust with zero-shot evaluation — where models attempt to reason about visual scenes with previously unseen entities (Suglia et al., 2020), which can be dangerous in practice for visually impaired people who need to rely on the response.

2.3 Existing Assistants for Sight Impaired People

There are a variety of commercial systems built to assist people with visual impairments. Some are applications, and others include specific hardware for the user. These systems fall into two categories: human-in-the-loop, or E2E.

Human-in-the-loop systems connect visually impaired people to a volunteer, or staff member, that is ready to answer visual questions. Examples include: *BeMyEyes*, *BeSpecular*, and *Aira* — with varying costs and wait times. These systems are very time-efficient thanks to the ability to have a *dialogue*, hence our focus on handling conversational utterances and follow-up questions. Human-in-the-loop systems also enable people with visual impairments to ask questions that involve artistic, cultural, or timely importance — like asking about Banksy's work — which E2E approaches cannot do (Fleet et al., 2020). There are several issues with this approach however, the major one being user *privacy*. The more affordable, and often free, services require untrained volunteers that receive images taken by visually impaired people. This is a huge security concern as the images may contain the user's name, address, medication, or children's photos in identifiable uniform (Fleet et al., 2020).

E2E systems, like *TapTapSee* and Microsoft *Seeing AI* are cloud services, so they do not have this privacy concern to the same extent. This concern becomes negligible if the system is open-source and can be set up at home, or keeps data on a device like *OrCam MyEye*. These E2E systems do also have their flaws however. They lack the mentioned ability to have a dialogue or understand culture, but more importantly, they cannot provide feedback on certainty. This makes it impossible to know whether the output is accurate. Similarly, it is very resource-intensive to tweak or extend E2E models, and a challenge to control specific behaviours (Samek et al., 2019). For example, it would be beneficial for the E2E system to be more cautious when answering questions about medication. This of course is a balance with human-in-the-loop systems that would answer accurately, but provide a stranger with medical information.

All of these systems offer general assistance to people with visual impairments, whereas we are concentrating on the kitchen domain. A system was recently developed which focused on improving the mealtime experience of visually impaired people after surveying them about their mealtime experiences (Chung et al., 2021). A virtual reality (VR)-based prototype was created to address a major issue that visually impaired people experienced: getting information about the location of food. The study also highlighted the anxiety faced by people with visual impairments about disturbing others for information or assistance. An automated system, such as Aye-saac, reduces the reliance on sighted people for kitchen and food-related tasks.

2.4 Natural Language Understanding

A dialogue system requires that an utterance, represented as text, is transformed into a meaningful representation for the Dialogue Manager (DM), thus enabling the formulation of a relevant response. In the context of conversational agents, this is known as NLU, which often refers to both identifying the intent of a user's input, and identifying which entities have been mentioned (Bunk et al., 2020). To reduce error-propagation between these sub-tasks, a multi-task architecture has been proposed called the Dual Intent Entity Transformer (DIET) classifier (Bunk et al., 2020). The joint modelling of the entity extraction and intent classification sub-tasks has been shown to improve performance, indicating that intent and entities closely interact with each other. Beyond improved accuracy, this model is faster to train than fine-tuning BERT for NLU.

Aye-saac relies on Rasa's implementation of the DIET classifier to achieve state-of-the-art NLU results. Rasa is a set of open-source libraries that can be used to create conversational agents (Bocklisch et al., 2017), and perform on par with paid NLU system, such as Microsoft's Language Understanding (LUIS) (Braun et al., 2017). Rasa offers the typical advantages of self-hosted open-source software such as adaptability and data control. This privacy is particularly important when interacting with visually impaired people in their homes.

3 Aye-saac

Aye-saac is a modular and extensible conversational VQA framework that is implemented as a collection of independent microservices. Isolating each service allows Aye-saac to utilise concurrency when analysing image data; ensuring more intensive operations do not hinder the system from responding to other requests. Figure 2 shows the flow of data between all the individual services, using RabbitMQ to control their communication.

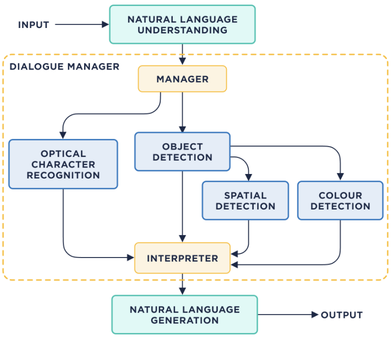


Figure 2: The microservices and data-flow in Aye-saac.

3.1 Suitability for the Kitchen

Some of Aye-saac's limitations within a kitchen environment were caused by its inability to detect common kitchen objects. Specifically, object detection was performed by a Single Shot Detector (SSD) with ResNet50 trained on the COCO dataset (Lin et al., 2020). The COCO dataset contains 80 object classes, of which we would only expect 30 to be commonly found in a kitchen. Examples of irrelevant classes include '*traffic light*' and '*giraffe*'.

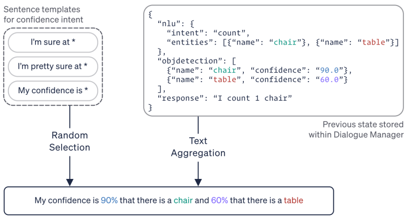
To improve Aye-saac's suitability in the kitchen, we combined the existing model with a baseline Faster R-CNN model, trained on the Epic-Kitchens-55 (EK) dataset (Damen et al., 2020). The latter model can identify 290 distinct objects that are commonly found in a kitchen. By combining both models Aye-saac is able to detect 299 kitchen-relevant object classes: 21 classes occur in both datasets, 9 are unique to COCO, and 269 are unique to EK. We retained the COCO model due to its superior accuracy (see Section 5.2).

3.2 Querying the Quantity

One issue for visually impaired users is knowing whether they have enough ingredients for a specific recipe (Gurari et al., 2018). Therefore, we introduced functionality to allow users to query the number of objects in a visual scene. The object detection only returns labels in singular form so we added a plural management feature to Aye-saac's NLU service. Therefore, "*How many eggs are there?*" can now be successfully answered and this plural management is extended across all existing intents related to object detection.

3.3 Quantifying Confidence

The detection of objects can be imprecise and it is critical to communicate this uncertainty to visually impaired users. For example, if the confidence regarding the number of eggs on the table is only 51%, the user could ask where the eggs are to manually count them. We therefore implemented a new intent that allows users to ask Aye-saac how confident it is about an object it detected. The response reports the image classification score from the object detection model as the confidence, as shown in Figure 3. It is important to note that the system could be 100% confident, but still be wrong.


Sentence templates for confidence intent
I'm sure at *
I'm pretty sure at *
My confidence is *
Previous state stored within Dialogue Manager
Random Selection
Text Aggregation

```
{
  "nlu": {
    "intent": "count",
    "entities": [{"name": "chair"}, {"name": "table"}]
  },
  "objdetection": [
    {"name": "chair", "confidence": "90.0"},
    {"name": "table", "confidence": "60.0"}
  ],
  "response": "I count 1 chair"
}
```

My confidence is 90% that there is a chair and 60% that there is a table

Figure 3: Illustration of how the NLG formulates a response, using the '*confidence*' intent as an example.

3.4 Relative Spatial Detection

Visually impaired people experience issues with locating food during mealtimes and, when given assistance options, prefer to have dish locations verbally described (Chung et al., 2021). We therefore developed spatial detection functionality to provide *usable* positioning information in relation to other objects. We created a list of 34 object classes taken from the COCO and EK datasets that we name '*anchors*' — corresponding to large items that are not expected to move very often, and thus the user will likely already know where they are, e.g. a kitchen sink or an oven. Additionally, we specify the expected relationships between an anchor and a query object from a list of prepositions; for example, a fridge has the spatial relationships '*in*', '*on*' and '*next to*' but not '*below*'. We prioritise positioning of query objects in relation to these anchors but in the absence of anchors, we attempt to give the position relative to people or hands that were found in the picture. Failing this, we return the absolute position of the object in the image.

4 Evaluation

4.1 Evaluation of Object Detection and VQA

We use a small sample of the VizWiz images and questions to evaluate the performance of the object detection and VQA (Gurari et al., 2018). Specifically, we used the VizWiz Dataset Browser (Bhattacharya and Gurari, 2019) to select images that are labelled as: suitable for object recognition, good quality, in the kitchen, with 10/10 confident answers. This gave a final evaluation set of 18 images with associated questions, asked by visually impaired people, and high confidence answers.

We used these images as input to the COCO and EK models and rated the resulting bounding boxes as correct or incorrect. We used the same images and questions to evaluate the VQA capabilities of Aye-saac and two E2E VQA models; Pythia (Jiang et al., 2018) and HieCoAtt (Lu et al., 2016). For all three systems, we compared the generated answers to the human provided answers.

4.2 Evaluation of Spatial Relationships

To evaluate the spatial relations, we first generated a small test set of image-question pairs from the GQA dataset (Hudson and Manning, 2019). Using the scene graphs from the training data, we gathered all the colour images that had: the word "*kitchen*" in at least one of their scene objects, the semantic type '*relation*', and began with the word "*where*". This gave a set of 18 image-question pairs which were asked to Aye-saac, Pythia, and HieCoAtt. The outputs were examined and rated as correct or incorrect and the *usability* of the spatial relationships were judged by the authors as usable or not in terms of whether the output could be used by a visually impaired user to locate the query object.

5 Results and Discussion

5.1 NLU

Following Rasa's evaluation guidelines, we generated the confusion matrix in Figure 4. Our NLU performs well but we can deduce that the intents '*identify*' and '*read text*' are mistaken several times. This can be improved with more training data.

5.2 Object Detection

Aye-saac relies on two pre-trained object detection models, the COCO model and the EK model. While the EK model is able to detect a much larger number of kitchen-specific objects than the COCO model, we found that the COCO model performed best on the image-question pairs detailed in Section 4.1. COCO identified 89.3% bounding boxes correctly, whereas the EK model identified 28.6%.

There is a need for an object detection model that is able to accurately detect a large number of kitchen-specific objects. Currently the baseline EK model has been used in Aye-saac because the better performing models have not been released publicly.

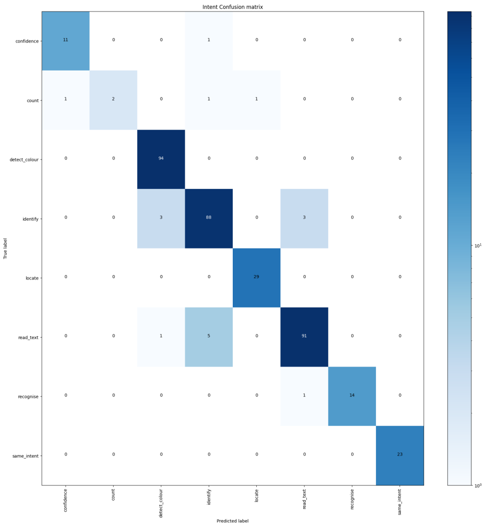


Figure 4: Intent confusion matrix comparing predicted intents against true labels.

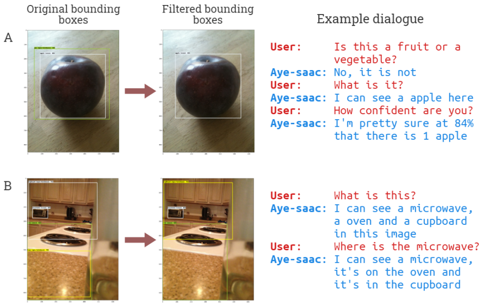
5.3 Spatial Relationships

Using the image-question pairs detailed in section 4.2 we tested how well Aye-saac determines the spatial relationships between queried objects. We rated the answers in terms of their correctness and usability. Overall, 77.8% of the answers were answered correctly and 66.7% of answers were deemed to be usable. 77.8% were answered using positioning relative to anchor objects, of which 71.4% were correct. Two of the questions were answered using positioning relative to people in the scene (100% answered correctly). The remaining questions were answered using absolute positioning in the image (100% answered correctly). The main reason for incorrect answers was multiple query bounding boxes resulting in stilted responses, e.g. "*I can see a sink and it's in the sink and a sink and it's left of the sink*". In one case the perspective of the objects was incorrectly interpreted, a microwave described as '*on*' the dining table instead of '*in front of*" the dining table.

We compared Aye-saac with the E2E systems, Pythia and HieCoAtt, see Table 1 for results. Both the E2E systems returned answers that were not usable, e.g. describing an objects position in relation to the "*counter*" or "*kitchen*". While technically the objects were indeed in a kitchen or on a counter, these answers do not help locate the query objects in the scene. Following this criterion, Pythia responded to 27.8% of the spatial relation questions

Table 1: Comparing accuracy and usability of Aye-saac spatial relationships versus E2E systems

System	Accuracy (%)	Usability (%)
Aye-saac	77.8	**66.7**
Pythia	**100.0**	27.8
HieCoAtt	50	5.6



User: Is this a fruit or a vegetable?
Aye-saac: No, it is not
User: What is it?
Aye-saac: I can see a apple here
User: How confident are you?
Aye-saac: I'm pretty sure at 84% that there is 1 apple

User: What is this?
Aye-saac: I can see a microwave, a oven and a cupboard in this image
User: Where is the microwave?
Aye-saac: I can see a microwave, it's on the oven and it's in the cupboard

Figure 5: Example Aye-saac interactions

usefully, all for microwave locations "*above stove*", and HieCoAtt only gave usable answers to 5.6% of the questions.

Aye-saac provides more detailed positions than the E2E systems by relating the position of objects to identified anchor points. The position detection could be further improved by accounting for multiple bounding boxes, e.g. in one case the location is reported in relation to four bounding boxes, one or two may be sufficient. Hardware changes could also improve the system, an addition of a Microsoft Kinect or stereo vision camera may help overcome depth perception issues. The camera could be placed at a high vantage point, but the object detection would then have trouble identifying objects as the EK model was trained on an egocentric dataset (Damen et al., 2020).

5.4 General VQA

Using the VizWiz image-question pairs detailed in section 4.1, we compared Aye-saac with two E2E systems — Pythia and HieCoAtt. When looking at accuracy alone, Pythia performed the best and answered 50% of the questions correctly, followed by HieCoAtt with 44%, and Aye-saac with 33%.

Figure 5A illustrates that Aye-saac currently suffers from a lack of deeper understanding. Here, a picture of an apple with the question "*Is this a fruit or a vegetable?*" is answered incorrectly as Aye-saac does not understand that an apple is a fruit. The E2E systems however are able to answer the question correctly. To enable this understanding, rule-based VQA systems like Aye-saac could be integrated with large Knowledge Bases (KBs) and ontologies on particular topics, and common sense knowledge. Some of these are very large and actively developed by communities of experts in the KBs respective domain. A few cross-domain examples include Wikidata (Vrandečić and Krötzsch, 2014), ConceptNet (Liu and Singh, 2004), and DBpedia (Auer et al., 2007); all part of the linked open data cloud (Auer et al., 2014).

A benefit of Aye-saac over the E2E systems is support for multi-turn dialogue so that users can query the confidence of given answers. Aye-saac achieves this by temporarily storing detected objects in its state. This feature could be expanded to support additional follow up questions, e.g. "*What is this?*" followed by "*and what colour is it?*". Aye-saac's modular design enables the addition of functionalities, like textual VQA (Ramil Brick et al., 2021), while E2E systems require retraining to cover additional objects or user queries.

Figure 5 illustrates another difficulty answering object positioning questions, determining the 3D position of objects with 2D images. Due to this, Aye-saac responds that the microwave is "*on the oven and in the cupboard*". Distinguishing between '*on*' and '*in*' is a complex semantic issue (Coventry et al., 2001; Richard-Bollans et al., 2020).

6 Conclusion

Blind and partially sighted people face many challenges in the kitchen. We presented our version of Aye-saac, a conversational VQA framework that aims to start tackling some of these challenges.

While our system can still be improved, we have shown that Aye-saac: (1) provides more *usable* responses than Pythia and HieCoAtt when asked for an object's location; (2) is transparent in design and also when detailing its own confidence of a previous response; (3) can have a multi-turn interaction; (4) still has an accurate NLU intent classifier with the added functionality and plural handling; and (5) can answer questions about 299 kitchen objects.

We had planned to run a user evaluation with visually impaired people in an accessible kitchen. Unfortunately due to COVID, we had to cancel this. In future work we could train a more accurate kitchen-specific object detection model on the EK dataset, integrate common-sense knowledge using KGs, and more. Once complete, we hope that a human evaluation could take place.

7 Acknowledgements

We couldn't have done this work without Julian Frabel, Noe Rivals, Guillaume Jeanne, Rachel Yu, Basile Lamarque, and Nejma Belkhanfar. The fourth author is funded by Wallscope and The Data Lab. Thank you all.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Sören Auer, Volha Bryl, and Sebastian Tramp. 2014. *Linked Open Data–Creating Knowledge Out of Interlinked Data: Results of the LOD2 Project*, volume 8661. Springer.

Nilavra Bhattacharya and Danna Gurari. 2019. VizWiz dataset browser: A tool for visualizing machine learning datasets. *arXiv*.

Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open source language understanding and dialogue management. *arXiv e-prints*, page arXiv:1712.05181.

Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P. Bigham. 2013. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2117–2126, New York, NY, USA. Association for Computing Machinery.

Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, page 174–185, Saarbrücken, Germany. Association for Computational Linguistics.

Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. DIET: Lightweight language understanding for dialogue systems. *arXiv e-prints*, page arXiv:2004.09936.

SeungA Chung, Soobin Park, Sohyeon Park, Kyungyeon Lee, and Uran Oh. 2021. Improving mealtime experiences of people with visual impairments. In *Proceedings of the 18th International Web for All Conference*, W4A '21, New York, NY, USA. Association for Computing Machinery.

Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of Memory and Language*, 44(3):376–398.

Dima Damen, Hazel Doughty, Giovanni Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2020. The EPIC-KITCHENS dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2135–213509.

Chancey Fleet, Cynthia Bennett, and Venkatesh Potluri. 2020. Vizwiz grand challenge workshop at cvpr 2020 - panel discussion with blind technology experts.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. IQA: Visual Question Answering in Interactive Environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4089–4098.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz grand challenge: Answering visual questions from blind people. *CoRR*, abs/1802.08218.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0.1: the winning entry to the VQA challenge 2018. *arXiv*, pages 4–6.

Nabila Jones, Hannah Elizabeth Bartlett, and Richard Cooke. 2019. An analysis of the impact of visual impairment on activities of daily living and vision-related quality of life in a visually impaired adult population. *British Journal of Visual Impairment*, 37(1):50–63.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73.

Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. 2018. Tell-and-answer: Towards explainable visual question answering using attributes and captions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1338–1346, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *arXiv preprint arXiv:1606.00061*.

Elisa Ramil Brick, Vanesa Caballero Alonso, Conor O'Brien, Sheron Tong, Emilie Tavernier, Amit Parekh, Angus Addlesee, and Oliver Lemon. 2021. Am i allergic to this? assisting sight impaired people in the kitchen.

Adam Richard-Bollans, Anthony Cohn, and Lucía Gómez Álvarez. 2020. Categorisation, typicality & object-specific features in spatial referring expressions. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, SpLU, pages 39–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Alessandro Suglia, Ioannis Konstas, Andrea Vanzo, Emanuele Bastianelli, Desmond Elliott, Stella Frank, and Oliver Lemon. 2020. CompGuessWhat?!: A Multi-task Evaluation Framework for Grounded Language Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7625–7641, Online. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yin Jiang Zhao, Yan Ling Li, and Min Lin. 2019. A review of the research on dialogue management of task-oriented systems. In *Journal of Physics: Conference Series*, volume 1267, page 012025. IOP Publishing.