Visual Question Answering and Beyond

My research goal is to develop Artificial Intelligence (AI) systems that can ‘see’ (i.e., understand the contents of an image: who, what, where, doing what?) and ‘talk’ (i.e., communicate the understanding to humans in free-form natural language). Applications of such systems include:

- Aiding visually impaired users in understanding their surroundings [9] (Human: ‘What is on the shelf above the microwave?’, AI: ‘Canned containers’),
- Aiding analysts in making decisions based on large quantities of surveillance data (Human: ‘What kind of car did the man in red shirt leave in?’, AI: ‘Blue Toyota Prius’),
- Teaching children through interactive demos (Kid: ‘What animal is that?’, AI: ‘That is Dall Sheep. You can find those in Alaska.’),
- Interacting with personal AI assistants (such as Alexa, Siri) (Human: ‘Is my laptop in my bedroom upstairs?’, AI: ‘Yes’, Human: ‘Is the charger plugged in?’),
- Making visual social media content more accessible (AI: ‘Your friend Bob just uploaded a picture from his Hawaii trip’, Human: ‘Great, is he at the beach?’, AI: ‘No, on a mountain’).

**VQA: Visual Question Answering.** As a first step towards building machines that can convey their understanding of visual content via natural language, in 2015 [8], together with my collaborators, I proposed free-form and open-ended Visual Question Answering (VQA). Given an image and a natural language question about the image (e.g., ‘What kind of store is this?’, ‘How many people are waiting in the queue?’, ‘Is it safe to cross the street?’), the machine’s task is to automatically produce an accurate natural language answer (‘bakery’, ‘5’, ‘yes’). Akin to a visual Turing test, answering any possible question about an image is one of the ‘holy grails’ of semantic scene understanding.

My colleagues and I collected and made publicly available (www.visualqa.org) a large scale dataset (>0.25M images, >0.76M questions, ∼10M answers) – first of it’s kind! To give a sense of scale, the dataset collection involved more than 10,000 Amazon Mechanical Turk workers. This is equivalent to more than 41,000 hours (∼4.7 years) for a single person working 24 hours a day!

Unlike existing computer vision tasks which either represent single narrowly-defined problem (e.g., image classification, activity recognition), or are difficult to evaluate (e.g., image captioning), the questions in our VQA dataset require a potentially vast set of AI capabilities to answer (see the image above) – fine-grained recognition (e.g., ‘What kind of cheese is on the pizza?’), object detection (e.g., ‘How many bikes are there?’), and commonsense (e.g., ‘Does this person have 20/20 vision?’). Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format.

Our work on VQA has witnessed tremendous interest in a short period of time (3 years) - over 900 citations of the papers [4,8], over 800 downloads of the dataset, and a best poster award at the Workshop on Object Understanding for Interaction (ICCV15). Most significantly, this work and our efforts since have resulted in the creation of a new multidisciplinary sub-field in AI and an entirely new line of research!

**VQA Challenge and Workshop.** In order to benchmark algorithms on the VQA dataset and
to push the state-of-the-art (SOTA) on VQA, I led the organization of the first VQA Challenge and the first VQA workshop at CVPR16. Both the challenge and the workshop were very well received. Approximately 25 teams from 8 countries across academia and industry participated in the challenge. Motivated by the success of the first VQA challenge and the workshop, I co-organized the second edition of the challenge and the workshop at CVPR17 which saw similar participation (24 teams) as the first year. Continuing the trend, I co-organized the third edition of the challenge and the workshop at CVPR18. This time, we saw a significant increase in the participation, with 40 teams participating in the challenge. It is exciting to see that the SOTA in VQA has improved by 17% in just 3 years - 2015 to 2018.

In addition to the improvement in the SOTA, the VQA dataset and the VQA Challenges led to the development of a variety of models and follow-up datasets proposed for this task. For example, we saw models using spatial attention [31], models that jointly reason about image and question attention [20], models which dynamically compose modules [6], models that study how to fuse the vision and language features [12], models that study how to combine top-down and bottom-up attention [5], and we have seen datasets as well as models that focus on visual reasoning and compositionality in language [2,3,17]. There have also been other tasks that built up on VQA, such as Visual Dialog [11].

**Analyzing the Behavior of Visual Question Answering Models.** After the release of our VQA dataset, a number of deep-learning models were proposed for VQA [6–8, 10, 12, 15, 16, 18–21, 23, 24, 26, 28–31, 33]. Curiously, the performance of most methods was clustered around 60-70% (compared to human performance of 83%) with a mere 5% gap between the top-9 entries on the VQA Challenge 2016. In order to identify the most fruitful directions for progress, in 2016, I developed novel techniques for characterizing the behavior of VQA models [1]. I analyzed several representative state-of-the-art VQA models [8,12,20], including the models developed by us [8] and presented three novel findings that expanded our understanding of VQA models at that time – despite the progress, the VQA models at that time were ‘myopic’ (tended to fail on sufficiently novel instances), often ‘jumped to conclusions’ (converged on a predicted answer after ‘listening’ to just half the question), and were ‘stubborn’ (did not not change their answers across images).

**Overcoming Priors for VQA.** Motivated by the findings of my previous work [1] (and work by others [14,17,32]) that VQA models were heavily driven by superficial correlations in the training data and lacked sufficient image grounding and compositionality, in my recent works, I addressed some of these issues by – 1) proposing new evaluation protocols [2,3], 2) a new model architecture [2], and 3) a novel objective function [22], as described below:

1) I proposed two new evaluation protocols for VQA – a) train and test sets have different prior distributions of answers for different question types [2], b) test question-answer pairs are compositionally novel compared to training question-answer pairs [3]. Specifically, I created two new splits of the VQA dataset [8] – a) Visual Question Answering under Changing Priors (VQA-CP), and b) Compositional VQA (C-VQA). I evaluated several existing VQA models on these new splits and found that their performance degrades significantly compared to the original VQA split. Thus, these proposed splits can serve as benchmarks to evaluate the degree of visual groundedness and compositionality in VQA models.
2) I also proposed a novel Grounded Visual Question Answering (GVQA) model [2] that contains inductive biases and restrictions in the architecture specifically designed to prevent the model from ‘cheating’ by primarily relying on priors in the training data. Specifically, GVQA explicitly disentangles the recognition of visual concepts present in the image from the identification of plausible answer space for a given question, enabling the model to more robustly generalize across different distributions of answers. GVQA significantly outperforms existing VQA models on VQA-CP. Above figure illustrates outputs from GVQA and that from an existing model (SAN) [31]. For the given questions in the test set, SAN predicts the prior answers for the respective question types (first few words of the question) from the training data, resulting in incorrect answers. However, GVQA, being more visually grounded than SAN, correctly answers the test questions.

3) Although GVQA can be built on top of any existing VQA model, it does require non-trivial changes in the architecture. In our NIPS18 paper [22], my colleagues and I proposed a simple drop-in regularizer that can be added to any existing VQA model by simply changing the objective function. In order to do this, we introduced a question-only model that takes the question encoding from the VQA model and must leverage language biases in order to predict the correct answer. We then pose training as an adversarial game between the VQA model and this question-only adversary – discouraging the VQA model from capturing language biases in its question encoding. This approach improved performance significantly for multiple base models (including GVQA), achieving state-of-the-art on VQA-CP.

Future Work: VQA and Beyond

VQA. Despite tremendous progress in VQA, there are some specific types of questions in VQA where the community has not made enough progress (mentioned below). In the short-term future, I want to develop fundamental techniques for adding new capabilities or skills to VQA models:

- **Counting**: The performance of state-of-art VQA models on counting questions (e.g., ‘How many people are standing in the queue?’, ‘How many slices of pizza are there?’) is only ∼45% (compared to the human performance of ∼83% and overall (across all questions) performance of state-of-art models of ∼71%). Clearly, the act of counting itself is not challenging – what is challenging is to parse the language, identify the referring expressions, grounding the referring expressions into visual concepts (e.g., detecting each individual slice of pizza, detecting each person who is standing and in the queue) – all of these are studied today as separate tasks. I will study unified models that include these components as modules.

- **Optical Character Recognition (OCR)**: Most of the SOTA VQA models are monolithic deep neural networks (without any specialized components). It is perhaps unsurprising why such models have not been able to make progress on OCR-type questions (e.g., ‘What does the street sign say?’, ‘What is the name of the building?’) – simply put, despite all strengths of deep learning, it seems implausible that the ability to convert pixels to text (OCR) will simply emerge in a monolithic network (i.e., the network will automatically learn to detect image regions containing text, convert pixels to text representations, realize when the question is about text, and insert that text into the answer) all from the error signal of downstream VQA accuracy.

Fortunately, we do not have to hope. OCR is a mature sub-field of computer vision. By utilizing the progress made in the OCR community, I want to develop VQA models with specifically designed OCR modules. I believe some of the fundamental challenges that we will have to address are – how to localize which entity in the image the question is about and how
to adapt existing OCR techniques to work with wide-variety of images in the VQA dataset.

- **Knowledge Based Reasoning:** We have not made much progress on questions that require knowledge based reasoning and common sense (e.g., ‘Does this person have 20/20 vision?’, ‘Is this food healthy?’). Such questions require an agent to understand what ‘20/20’ vision means and what types of food items are healthy, in addition to understanding the visual content – recognizing that the person is wearing spectacles and that it is a fast food item. I believe a limiting factor in this area is lack of existence of a large-scale and open-ended Knowledge Based (KB) VQA dataset. The only existing dataset on KB-VQA is the fact-based VQA dataset [25]. However, this dataset consists of only ∼6000 questions, and a closed set of visual concepts and KB facts. Utilizing my experiences from VQA dataset collection, I plan on creating a much larger and open-ended KB-VQA dataset, in order to push the progress in this direction.

**Beyond: From Vision and Language to Actions.** Most of my past work has been towards building agents that can ‘see’ and ‘talk’. However, for a lot of practical applications (e.g., physical agents navigating inside our houses executing natural language commands) we need agents that can not only ‘see’ and ‘talk’ but can also take actions. In the medium-term and long-term future, I want to generalize vision and language agents to be able to take actions.

In this space of building agents that can ‘see’, ‘talk’ and ‘act’, one bold initiative which was well ahead of its time was the SHRDLU [27] project, studied by Terry Winograd in 1972 – an agent that operates on a table top scene consisting of several blocks such as cuboid, cone, containers etc.; there is a teacher which instructs the agent what to do (e.g., ‘pick up a red block’) and the agent either executes an action or asks a question (if the instruction is not clear) (e.g., ‘By “It” I assume you mean the block which is taller than the one I am holding.’) (see the figure on right).

However, SHRDLU was a hand-engineered rule-based system. I want to build learning based SHRDLU agent so that it can scale to richer language / realistic scenes / more complicated actions.

As a first step in this direction, in an ongoing project, I am working on the following - how can we train agents to follow language instructions grounded in visual data (e.g., ‘Add a red sphere’, ‘Add a large cylinder’) and execute actions to generate scenes that are consistent with the given instruction. Using reinforced adversarial learning framework [13], I have taken the first step towards training agents that can follow simple instructions (as mentioned above).

In an effort to move closer to SHRDLU like agent, in the **medium-term future**, I plan on building on top of the above set-up, so as to generalize it from creating scenes to actually moving objects, albeit in a simulated physical environment. Working with such a setup poses multiple challenges – how to work with continuous control, how to reason about physics, etc. I also plan on using more complicated instructions such as those involving multiple objects and relationships (e.g., “Add a red sphere to the right of the big cylinder.”). The presence of multiple objects in a scene makes the problem of grounding language into vision and actions even more challenging.

In the **long-term future**, my goal is to build a SHRDLU like agent – moving from instruction following to engaging in a dialog with a human user and grounding that into vision and actions. To
Aishwarya Agrawal

Research Statement

achieve this, I plan on collecting natural language dialog data from humans, and using that to pre-train the agent using Supervised Learning, so that the agent can learn the vocabulary and language model of how humans speak. I then plan on fine-tuning the agent using Reinforcement Learning so that it can learn a policy for engaging in dialog and executing natural language commands.

Building such an agent will be a step towards building physical agents that can assist humans in day-to-day tasks (e.g., Human: ‘What time is it now?’, AI: ‘5pm’, Human: ‘It’s time to water the plants. Can you please do that?’), because engaging in a dialog with a human user and grounding that into vision and actions is an essential component for such physical agents.

References