Introduction
QA as Standardized Tests for Machine Reasoning

• **Question Answering** = computer systems that automatically answer natural language questions about knowledge by humans.
  - Not simple search-and-retrieve.
    - E.g. “what affects her mobility?”

• **Why question answering?**
  - Humans learn by answering questions.
  - QA can be used to formulate other tasks
    - E.g. “what is present in the image?” (recognition), “what action has the person in the video performed?”

*Q: “What affects her mobility?”*
Learning to Reason in QA form

• Input:
  • A knowledge context \( C \)
  • A query \( q \)

• Output: an answer satisfying

\[
\tilde{a} = \arg\max_{a \in A} P_\theta(a \mid q, C)
\]

• \( C \) can be
  • structured: knowledge graphs
  • unstructured: text, image, sound, video

Q: Is it simply an optimization problem like recognition, detection or even translation?
→ No, because the logics from \( C, q \) into \( a \) is more complex than other solved optimization problems
→ We can solve (some parts of) it with good structures and inference strategies

Q: “What affects her mobility?”

[Slide credit: Vuong Le]
Overall Architecture of General QA

- Question representation
- Context representation (domain specific)
- Context-question interaction (reasoning)
- Answer decoder
Lecture 7: Textual QA
Private schools, also known as independent schools, non-governmental, or nonstate schools, [...] thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, [...].

Q: Rather than taxation, what are private schools largely funded by?

A: charging their students tuition

Size: 151,054 samples
Task: given context information as a paragraph, predict the text span contains the correct answer.

Other Tasks

Cloze

**P:** You will need 3/4 cup of black berries ... Pour the mixture into cups and insert a popsicle stick in it or pour it in a popsicle maker. Place the cup ... in the freezer. ...

**Q:** Choose the best title for the missing blank to correctly complete the recipe. *Ingredients, __, Freeze, Enjoying*

**Candidates:** (A) Cereal Milk Ice Cream (B) Ingredients (C) Pouring (D) Oven

**Answer:** C

Multiple choice

**P:** It was Jessie Bear’s birthday. She ...

**Q:** Who was having a birthday?

**Candidates:** (A) Jessie Bear (B) no one (C) Lion (D) Tiger

**Answer:** A

Open-ended

**P:** ...Mark decides to broadcast his final message as himself. They finally drive up to the crowd of protesting students, .... The police step in and arrest Mark and Nora....

**Q:** What are the students doing when Mark and Nora drive up?

**Answer:** Protesting
Machine Comprehension Test
Word Representations

Word2Vec

GloVe
Word Representations in Context

• Word representations should vary depending on context
Recurrent Neural Networks

- Variants: LSTM, GRU
- Advantages:
  - Good for sentences/short text
  - Robust in practice
- Disadvantages:
  - Slow, computational costly
  - Cannot parallelize
  - Not good for very long sequences.

\[ y_t = f(x_t, y_{t-1}) \]
Self-Attention

• Advantages:
  • Good at capturing long range dependencies
  • Can capture co-reference chains
  • Parallelizable and fast

• Disadvantages:
  • Memory intensive
  • Hyper-parameters tuning

- K: attention heads
- T: sequence length
- $\alpha_{j,i}$: self-attention weights

$$\tilde{x}_t^k = \sum_{j=1}^{T} \alpha_{j,i}^k x_j$$
$$y_t = f(\tilde{x}_t^1, \ldots, \tilde{x}_t^K)$$
Reasoning Approaches in MRC

Supervised training

- Attention-based encoders
  - Passage of text
  - Question
  - Answer

QA as a downstream task of unsupervised pretrained models

- Transformer-based models
  - Large text corpus
  - Transfer

- Neural net encoder for QA
  - Passage of text
  - Question
  - Answer
Bi-Directional Attention Flow Model

BERT: Transformer That Predicts Its Own Masked Parts

BERT is like parallel approximate pseudo-likelihood

- ~ Maximizing the conditional likelihood of some variables given the rest.

- When the number of variables is large, this converges to MLE (maximum likelihood estimate).

https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

[Slide credit: Truyen Tran]
BERT – Fine-tuning for MRC

(c) Question Answering Tasks: SQuAD v1.1

<table>
<thead>
<tr>
<th>System</th>
<th>Dev EM</th>
<th>Dev F1</th>
<th>Test EM</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Leaderboard (Oct 8th, 2018)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>-</td>
<td>-</td>
<td>82.3</td>
<td>91.2</td>
</tr>
<tr>
<td>#1 Ensemble - nlnet</td>
<td>-</td>
<td>-</td>
<td>86.0</td>
<td>91.7</td>
</tr>
<tr>
<td>#2 Ensemble - QANet</td>
<td>-</td>
<td>-</td>
<td>84.5</td>
<td>90.5</td>
</tr>
<tr>
<td>#1 Single - nlnet</td>
<td>-</td>
<td>-</td>
<td>83.5</td>
<td>90.1</td>
</tr>
<tr>
<td>#2 Single - QANet</td>
<td>-</td>
<td>-</td>
<td>82.5</td>
<td>89.3</td>
</tr>
<tr>
<td><strong>Published</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiDAF+ELMo (Single)</td>
<td>-</td>
<td>-</td>
<td>85.8</td>
<td>-</td>
</tr>
<tr>
<td>R.M. Reader (Single)</td>
<td>78.9</td>
<td>86.3</td>
<td>79.5</td>
<td>86.6</td>
</tr>
<tr>
<td>R.M. Reader (Ensemble)</td>
<td>81.2</td>
<td>87.9</td>
<td>82.3</td>
<td>88.5</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT\textsubscript{BASE} (Single)</td>
<td>80.8</td>
<td>88.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BERT\textsubscript{LARGE} (Single)</td>
<td>84.1</td>
<td>90.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BERT\textsubscript{LARGE} (Ensemble)</td>
<td>85.8</td>
<td>91.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BERT\textsubscript{LARGE} (Sgl.+TriviaQA)</td>
<td>84.2</td>
<td>91.1</td>
<td>85.1</td>
<td>91.8</td>
</tr>
<tr>
<td>BERT\textsubscript{LARGE} (Ens.+TriviaQA)</td>
<td>86.2</td>
<td>92.2</td>
<td>87.4</td>
<td>93.2</td>
</tr>
</tbody>
</table>

Figures from Devlin et al. 18'
Lecture 8: Image Question Answering
Recall the Learning to Reason formulation

- **Input:**
  - Context $C$ given by an **image**
  - A query $q$

- **Output:** an answer satisfying
  \[
  \hat{a} = \arg\max_{a \in \mathcal{A}} P_\theta(a \mid q, C)
  \]

**Q:** “What affects her mobility?”
Why VQA is an AI testbed?

**Question:** What can the red object on the ground be used for?

**Answer:** Fighting fires.

**Support Fact:** Fire hydrants can be used for fighting fires.

VQA

- **Computer Vision (1)**
- **Natural Language Processing (2)**
- **Machine Learning (3)**
- **Reasoning (4)**
Why VQA is an AI testbed?

Adapted from [Somak et al., 2019]
Applications of VQA

- Aid visually-impaired users

Image credit: ARIA
Applications of VQA

- Surveillance and visual data summarization

*What did the man in red shirt do before entering the building?*

Image credit: journalistsresource.org

shutterstock.com • 289173068
VQA: Question types

**Open-ended**
- Is this a vegetarian pizza?
- What is the red thing in the photo?

**Multi-choice**
(Q) What is the red thing in the photo?
(A)  (1) capsicum   (2) beef
     (3) mushroom   (4) cheese

**Counting**
- How many slices of pizza are there?

(VQA, Agrawal et al., 2015)
VQA: Image QA datasets

(Q) What is in the picture?
(Q) Is this a vegetarian pizza?

(Q) What is the brown animal sitting inside of?
(Q) Is there a bag to the right of the green door?

(Q) How many objects are either small cylinders or metal things?
(Q) Are there an equal number of large things and metal spheres?
Dual-view System for VQA

System 1: Intuitive
- Fast
- Implicit/automatic
- Pattern recognition
- Multiple

System 2: Analytical
- Slow
- Deliberate/rational
- Careful analysis
- Single, sequential

(Deep feature extraction, e.g. CNN, GloVe/BERT etc.)
(Cross-modality interaction)
Attention-based VQA Methods

• Unidirectional attention
  • Find relation score between parts in the context C to the question q:

\[
s_i = f(q, \mathbf{w}_j^c)
\]

Options for f:
  • \(s_i = \tanh(W^c_i \mathbf{w}_i^c + W^q q)\) Hermann et al. (2015)
  • \(s_i = q^T W^s \mathbf{w}_i^c\) Chen et al. (2016)

• Normalized by softmax into attention weights

\[
\alpha_i = \frac{\exp(W s_i)}{\sum_j \exp(W s_j)}
\]

• Attended context vector:

\[
i = \sum_i \alpha_i \mathbf{w}_i^c
\]

\(\rightarrow\) We can now extract information from the context that is “relevant” to the query
Bottom-up-top-down attention (Anderson et al 2017)

• Bottom-up set construction: Choosing Faster-RCNN regions with high class scores
• Top-down attention: Attending on visual features by question

→ Q: How about attention from vision objects to linguistic objects?

[Slide credit: Vuong Le]
Bi-directional attention

• Question-context similarity measure
  \[ s_i = f(q, w^c_j) \]

• Question-guided context attention
  • Softmax across columns

• Context-guided question attention
  • Softmax across rows

→ Q: Probably not working for image QA where single words does not have the co-reference with a region?
Hierarchical co-attention for ImageQA

- The co-attention is found on a word-phrase-sentence hierarchy
  → better cross-domain co-references

→ Q: Can this be done on text QA as well?
→ Q: How about questions with many reasoning hops?
Compositional Reasoning: Why do we care?

- Visual data and text data are compositional by nature.
- **Principle of compositionality**: “the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them”

What color is the thing with the same size as the cyan cylinder?

- The network guessed the most common color in the image.
- Linguistic bias.
- Requires **multi-step reasoning**: find cyan cylinder → locate another object of the same size → determine its color (green).

Figure from CLEVR dataset
Multi-step Compositional Reasoning

• Complex question need multiple hops of reasoning
• Relations inside the context are multi-step themselves
• Single shot of attention won’t be enough
• Single shot of information gathering is definitely not enough

→ Q: How to do multi-hop attentional reasoning?
Multi-step reasoning - Memory, Attention, and Composition

- Attention reasoning is done through multiple sequential steps.
- Each step is done with a recurrent neural cell
- **What is the key differences to the normal RNN (LSTM/GRU) cell?**
  - Not a sequential input, it is sequential processing on static input set.
  - Guided by the question through a controller.
Multi-step Attentional Reasoning

• At each step, the controller decide what to look next
• After each step, a piece of information is gathered, represented through the attention map on question words and visual objects
• A common memory kept all the information extracted toward an answer
Multi-step Attentional Reasoning

- Step 1: attends to the "tiny blue block", updating $\textbf{m}_1$
- Step 2: look for "the sphere in front" $\textbf{m}_2$.
- Step 3: traverse from the cyan ball to the final objective – the purple cylinder,
From Spatial Reasoning to Object-centric Reasoning

• Grid representation is irrespective of the fine-grained semantics of images.
• Region proposals are of the same semantic abstract with words -> help visual grounding.
• Interpretability.

Dynamic Reasoning Graphs

• On complex questions, multiple sets of relations are needed
• We need not only multi-step but also multi-form structures
• Let’s do multiple dynamically-built graphs!
The questions so far act as an unstructured command in the process

Aren’t their structures and relations important too?
Cross-modality Graph Interactions for VQA

- Two types of nodes: Linguistic entities and visual objects
- Two types of edges:
  - Visual
  - Linguistic-visual binding (*as a fuzzy grounding*)
- Adaptively updated during reasoning

LOGNet, T.M Le et.al. IJCAI2020
Language-binding Object Graph (LOG) Unit

- Graph constructor: build the dynamic vision graph
- Language binding constructor: find the dynamic L-V relations
LOGNet: multi-step visual-linguistic binding

- Object-centric representation ✓
- Multi-step/multi-structure compositional reasoning ✓
- Linguistic-vision detail interaction ✓
Dynamic Language-Vision Graphs in Actions

Question: Is the color of the big matte object the same as the large metal cube?
Prediction: yes  Answer: yes

Question: There is a tiny purple rubber thing; does it have the same shape as the brown object that is on the left side of the rubber sphere?
Prediction: no  Answer: no
Reasoning as Query-driven Program

- Reasoning as laying out modules to reach an answer
- Composable neural architecture $\rightarrow$ question parsed as program (layout of modules)
- A module is a function $(x \rightarrow y)$, could be a sub-reasoning process $((x, q) \rightarrow y)$.

What Do the Modules Learn?

**question:** do the small cylinder that is in front of the small green thing and the object right of the green cylinder have the same material? 

**ground-truth answer:** no

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="layout1.png" alt="Layout" /></td>
<td><img src="find1.png" alt="Find" /></td>
<td><img src="relocate1.png" alt="Relocate" /></td>
<td><img src="filter1.png" alt="Filter" /></td>
<td><img src="find2.png" alt="Find" /></td>
<td><img src="relocate2.png" alt="Relocate" /></td>
<td><img src="compare1.png" alt="Compare" /></td>
</tr>
<tr>
<td>&quot;yes&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="layout2.png" alt="Layout" /></td>
<td><img src="find2.png" alt="Find" /></td>
<td><img src="relocate2.png" alt="Relocate" /></td>
<td><img src="filter2.png" alt="Filter" /></td>
<td><img src="find3.png" alt="Find" /></td>
<td><img src="relocate3.png" alt="Relocate" /></td>
<td><img src="filter3.png" alt="Filter" /></td>
<td><img src="compare2.png" alt="Compare" /></td>
</tr>
<tr>
<td>&quot;no&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Current Trend in VQA: Learning to reason with less labels

• Data augmentation with analogical and counterfactual examples
• Question generation
• Self-supervised learning for question answering
• Learning with external knowledge graphs
Data Augmentation with Analogical and Counterfactual Examples

- **Poor generalization** when training under independent and identically distributed assumption.
- **Intuition**: augmenting counterfactual samples to allow machines to understand the critical changes in the input that lead to changes in the answer space.
  - Perceptually similar, yet
  - Semantically dissimilar realistic samples

Visual counterfactual example

Language counterfactual examples

Question Generations

- Question answering is a few-shot learning problem. Question generation helps cover a wider range of concepts.

- Question generation can be done with either supervised and unsupervised learning.
Visual QA as a Down-stream Task of Visual-Language BERT Pre-trained Models

Numerous pre-trained visual language models during 2019-2021.

- **Single-stream model**
  - Cross-modal Transformer
  - VisualBERT (Li, Liunian Harold, et al., 2019)
  - VL-BERT (Su, Weijie, et al., 2019)
  - UNITER (Chen, Yen-Chun, et al., 2019)
  - 12-in-1 (Lu, Jiasen, et al., 2020)
  - Pixel-BERT (Huang, Zhicheng, et al., 2019)
  - OSCAR (Li, Xiujun, et al., 2020)

- **Two-stream model**
  - Cross-modal Transformer
  - Multi-layer Transformer
  - ViLBERT (Lu, Jiasen, et al., 2019)
  - LXMERT (Tan, Hao, and Mohit Bansal, 2019)

[Slide credit: Licheng Yu et al.]
Learning with External Knowledge

Why external knowledge for reasoning?

- Questions can be beyond visual recognition (e.g. firetrucks usually use a fire hydrant).
- Human’s prior knowledge for cognition-level reasoning (e.g. human’s goals, intents etc.)

Q: What sort of vehicle uses this item?  
A: firetruck

Q: What is the sports position of the man in the orange shirt?  
A: goalie/goalkeeper


Learning with External Knowledge


Lecture 9: Video/Movie Question Answering
Recall the Learning to Reason formulation

• Input:
  • Context $C$ is a dynamic scene
  • A query $q$

• Output: an answer satisfying

$$\tilde{a} = \arg\max_{a \in A} P_\theta(a \mid q, C)$$

Q: What does the boy with a brown hoodie do before running away? A: *flip to the front side*
Challenges

• Difficulties in data annotation.
• Content for performing reasoning spreads over space-time and multiple modalities (videos, subtitles, speech etc.)
Video QA Datasets

Movie QA
(Tapaswi, M., et al., 2016)

MSRVTT-QA and
MSVD-QA
(Xu, D., et al., 2017)

TGIF-QA
(Jang, Y., et al., 2017)

TVQA/TVQA+
(Lei, J., et al., 2018)

CLEVRER
(Yi, K., et al., 2019)

KnowIT VQA
(Garcia, N., et al., 2020)
Video QA datasets

(TGIF-QA, Jang et al., 2018)

Q: What does the man do 5 times?
A: (0) step
   (2) sway head
   (5): move body to the front

Q: What does the man do before turning body to left?
A: (0) run a cross a ring
   (2) pick up the man’s hand
   (5): breath

(CLEVRER, Yi, Kexin, et al., 2020)

Q: What color is the last object to collide with the green cube?
A: cyan

Q: Which of the following is responsible for the collision between the metal cube and the cylinder?
A: (a) The presence of the brown rubber cube
   (b) The sphere's colliding with the cylinder
   (c) The rubber cube's entrance
   (d) The collision between the metal cube and the sphere
Video QA as a spatio-temporal extension of Image QA

(a) Extended end-to-end memory network
(b) Extended simple VQA model
(c) Extended temporal attention model
(d) Extended sequence-to-sequence model

Spatio-temporal cross-modality alignment

Key ideas:

• Explore the correlation between vision and language via attention mechanisms.

• Joint representations are query-driven spatio-temporal features of a given videos.

Memory-based Video QA

Key ideas:
• DMN refines attention over a set of facts to extract reasoning clues.
• Motion and appearance features are complementary clues for question answering.

Memory-based Video QA

Key differences:
• Learning a joint representation of multimodal inputs at each memory read/write step.
• Utilizing external question memory to model context-dependent question words.

Multimodal reasoning units for Video QA

- **CRN:** Conditional Relation Networks.
- **Inputs:**
  - Frame-based appearance features
  - Motion features
  - Query features
- **Outputs:**
  - Joint representations encoding temporal relations, motion, query.

Object-oriented spatio-temporal reasoning for Video QA

- **OSTR**: Object-oriented Spatio-Temporal Reasoning.

- **Inputs**:
  - Object lives tracked through time.
  - Context (motion).
  - Query features.

- **Outputs**:
  - Joint representations encoding temporal relations, motion, query.

---

Video QA as a down-stream task of video language pre-training
VideoBERT: a joint model for video and language representation learning

• Data for training: Sample videos and texts from YouCook II.

VideoBERT: a joint model for video and language representation learning

Pre-training

- Linguistic representations:
  - Tokenized texts into WordPieces, similar as BERT.

- Visual representations:
  - S3D features for each segmented video clips.
  - Tokenized into clusters using hierarchical k-means.

VideoBERT: a joint model for video and language representation learning

CLIPBERT: video language pre-training with sparse sampling

Procedure:
• Pretraining on large-scale image-text datasets.
• Finetuning on video-text tasks.

From short-form Video QA to Movie QA

Subtitle:

00:00:0.395 --> 00:00:1.896  
(Keith:) I’m not gonna stand here and let you accuse me

00:00:1.897 --> 00:00:4.210  
(Keith:) of killing one of my best friends, all right?

00:00:8.851 --> 00:00:10.394  
(Castle:) You hear that sound?

Question: What did Keith do when he was on the stage?

Choice 1: Keith drank beer  
Choice 2: Keith played drum  
Choice 3: Keith sing to the microphone  
Choice 4: Keith played guitar  
Choice 5: Keith got off the stage and walked out

Baseline: Keith played guitar  
HCRN: Keith got off the stage and walked out  
Ground truth: Keith got off the stage and walked out
Conventional methods for Movie QA

Question-driven multi-stream models:
• Short-term temporal relationships are less important.
• Long-term temporal relationships and multimodal interactions are key.
• Language is dominant over visual counterpart.


**HERO: large-scale pre-training for Movie QA**

- Pre-trained on 7.6M videos and associated subtitles.
- Achieved state-of-the-art results on all datasets.

<table>
<thead>
<tr>
<th>Method \ Task</th>
<th>TVR R@1</th>
<th>TVR R@10</th>
<th>TVR R@100</th>
<th>How2R R@1</th>
<th>How2R R@10</th>
<th>How2R R@100</th>
<th>TVQA Acc.</th>
<th>TVQA Acc.</th>
<th>TVQA Acc.</th>
<th>VIOLIN Acc.</th>
<th>VIOLIN Acc.</th>
<th>VIOLIN Acc.</th>
<th>TVC Bleu</th>
<th>TVC Rouge-L</th>
<th>TVC Meteor</th>
<th>TVC Cider</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA Baseline</td>
<td>3.25</td>
<td>13.41</td>
<td>30.52</td>
<td>2.06</td>
<td>8.96</td>
<td>13.27</td>
<td>70.23</td>
<td>-</td>
<td>67.84</td>
<td>10.87</td>
<td>32.81</td>
<td>16.91</td>
<td>45.38</td>
<td>69.12</td>
<td>65.20</td>
<td>34.11</td>
</tr>
<tr>
<td>HERO</td>
<td>6.21</td>
<td>19.34</td>
<td>36.66</td>
<td>3.85</td>
<td>12.73</td>
<td>21.06</td>
<td>73.61</td>
<td>73.81</td>
<td>68.59</td>
<td>12.35</td>
<td>34.16</td>
<td>17.64</td>
<td>49.98</td>
<td>72.50</td>
<td>68.91</td>
<td>35.28</td>
</tr>
</tbody>
</table>
End of Lecture 7+8+9

https://neuralreasoning.github.io