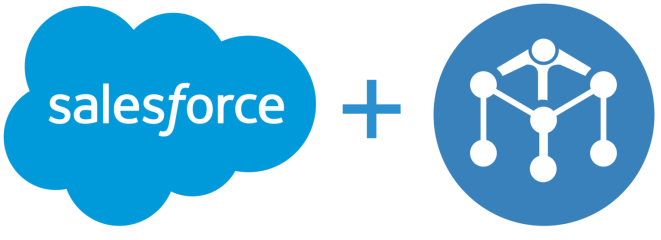

# Dynamic memory networks for visual and textual question answering

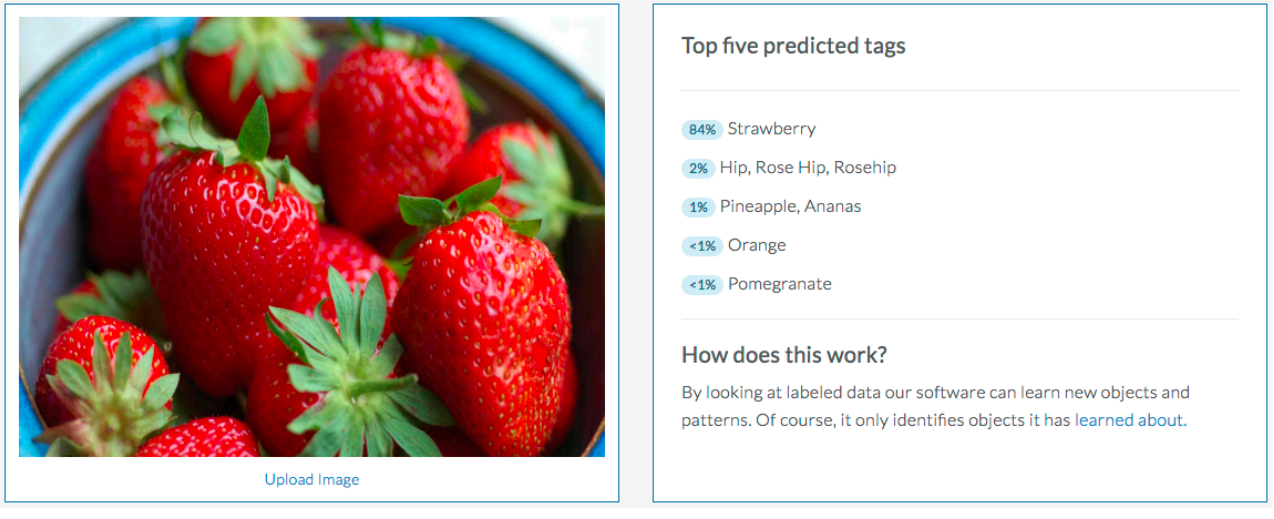
Stephen Merity (@smerity)




Joint work with the MetaMind team:
Caiming Xiong, Richard Socher, and more

# Classification

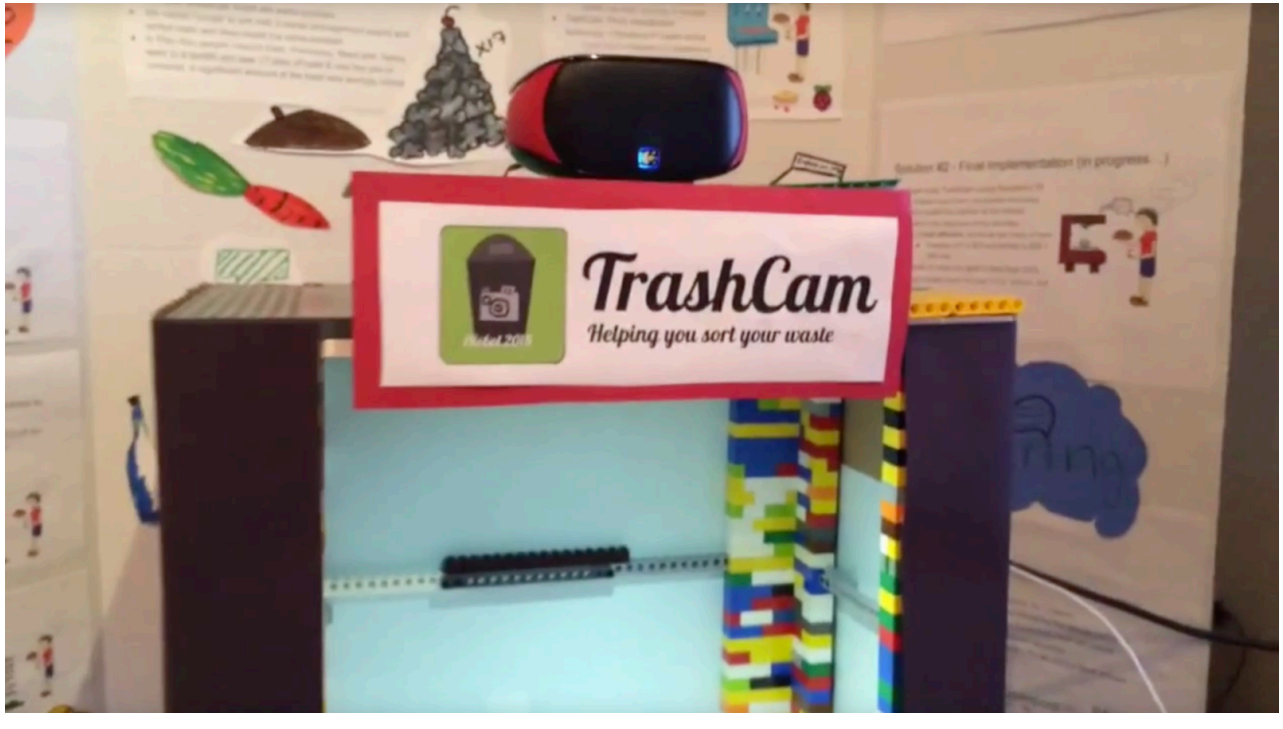
With good data, deep learning can give high accuracy in image and text classification




It's trivially easy to train your own classifier with near zero ML knowledge

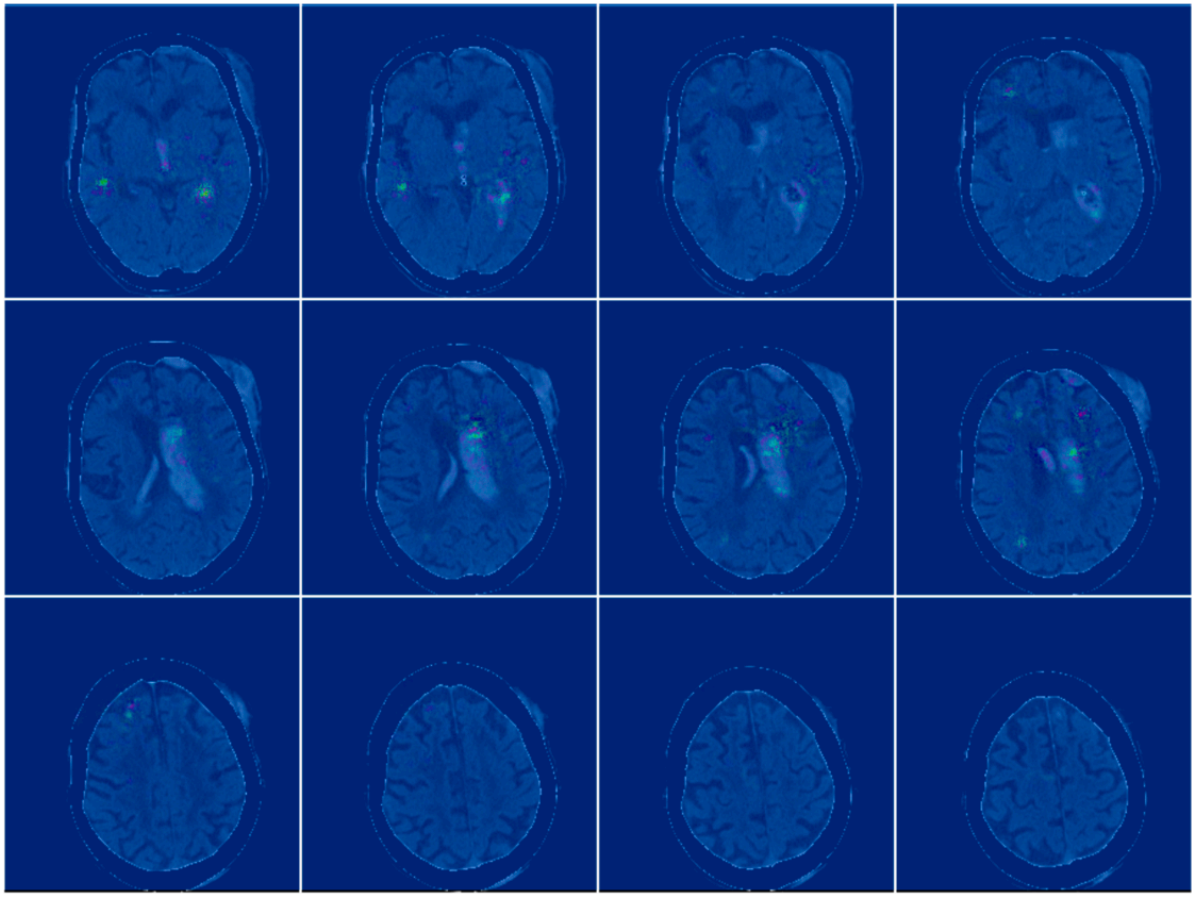
# It's so easy that ...

6th and 7th grade high school students created a custom vision classifier for *TrashCam*



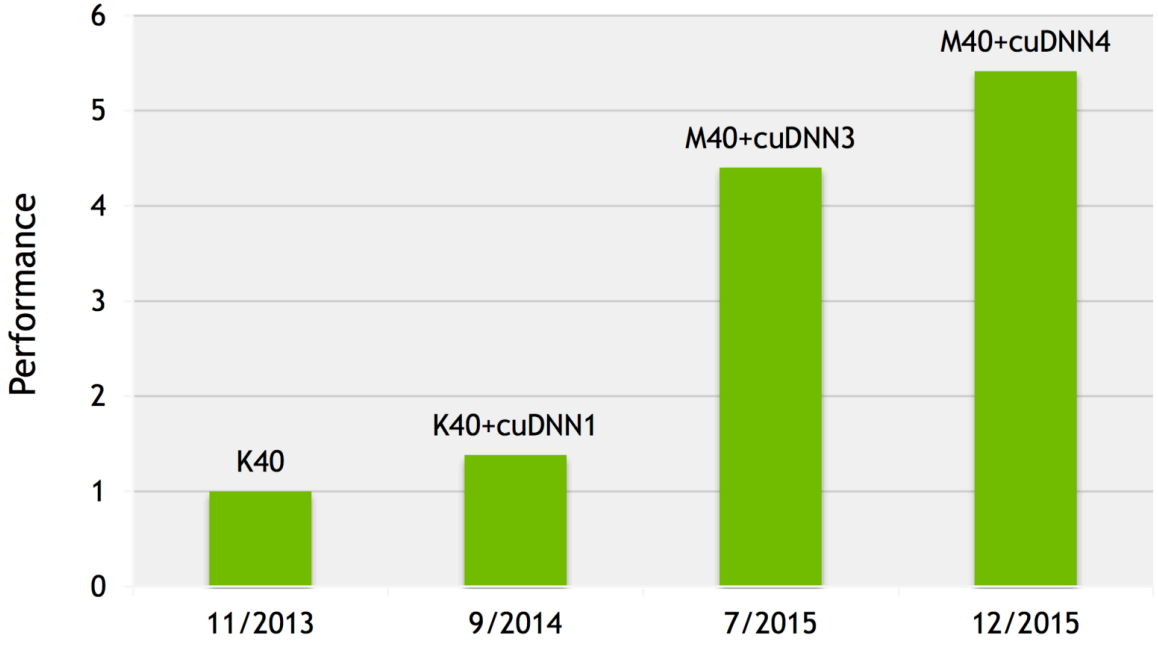

[*Trash*, *Recycle*, *Compost*] with 90% accuracy

# Intracranial Hemorrhage



Work by MM colleagues: Caiming Xiong, Kai Sheng Tai, Ivo Mihov, ...

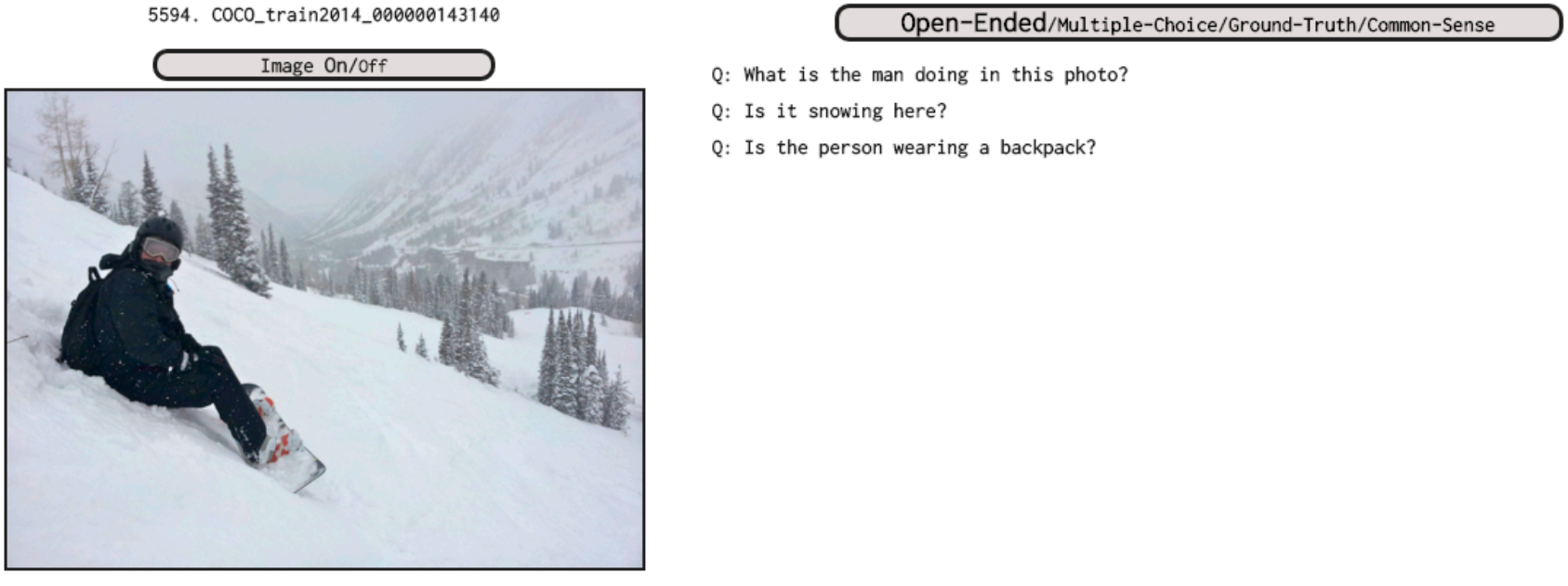
# Advances leveraged via GPUs




AlexNet training throughput based on 20 iterations

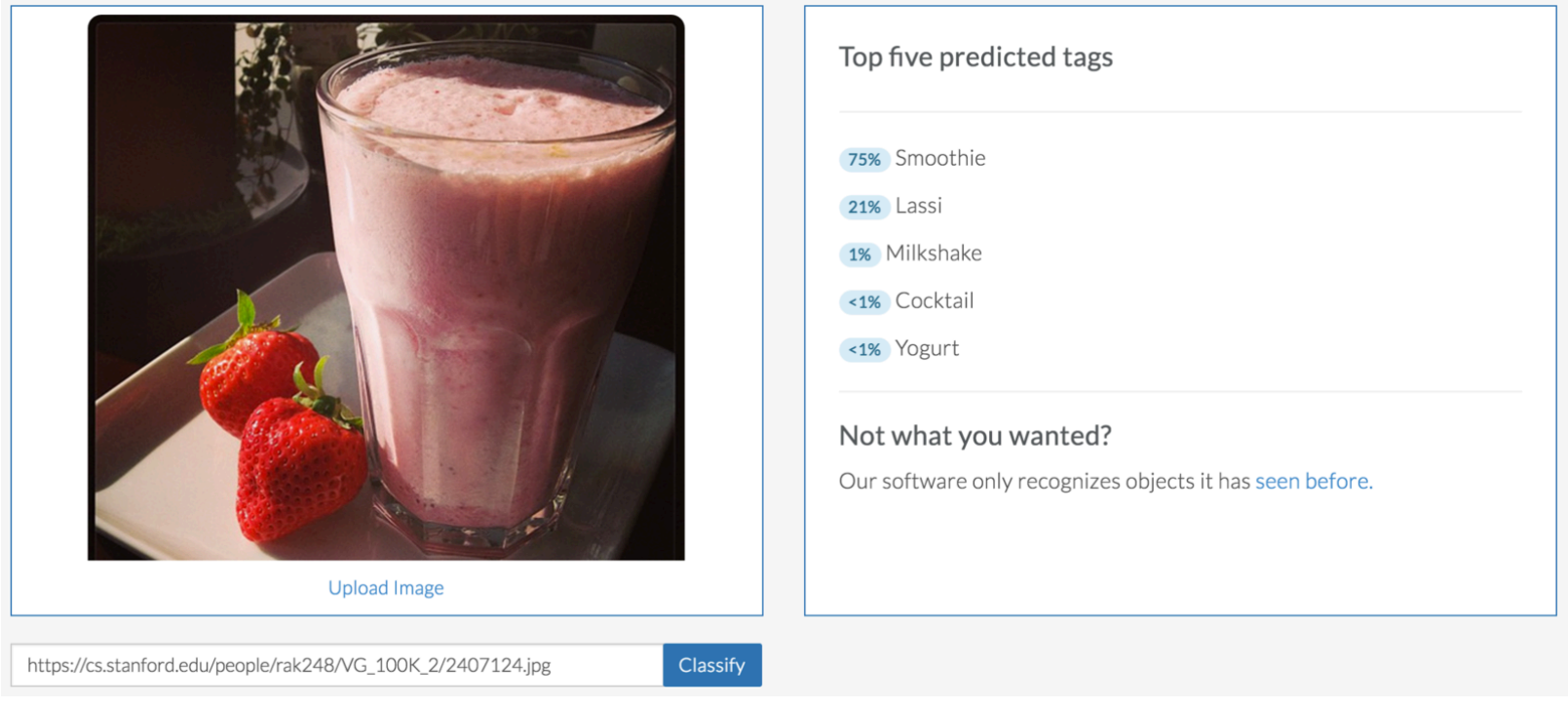
Slide from Julie Bernauer's NVIDIA presentation

# Beyond classification …




VQA dataset: http://visualqa.org/

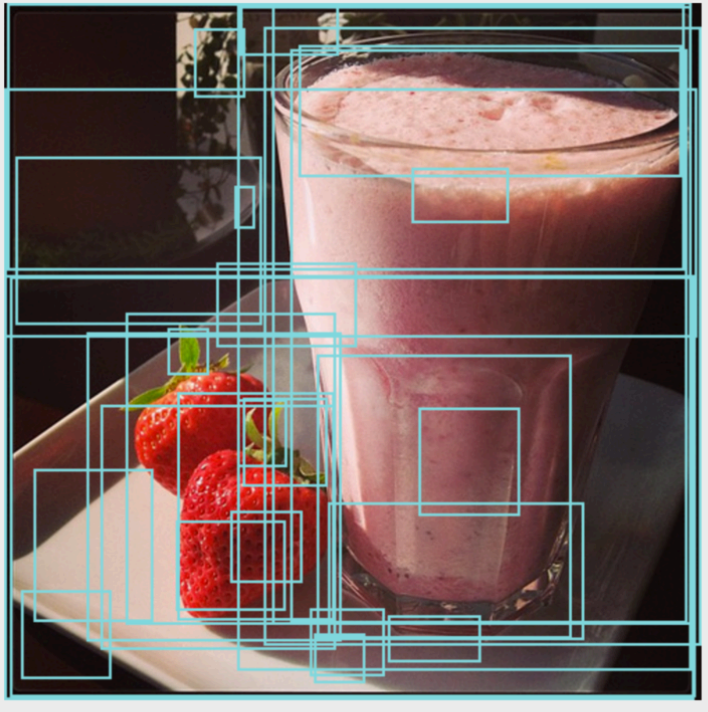
# Beyond classification …




* TIL Lassi = popular, traditional, yogurt based drink from the Indian Subcontinent

# Question Answering

| Regions | Attributes | Relationships |
|---|---|---|
| part of a glass | strawberries is lucious | part of glass |
| inner part of a tray | tray is white | inner part of inner part |
| part of a strawberry | glass is big | part of strawberry |
| base of a glass | towel is flowered | base of glass |
| strawberry level on the glass | strawberry top is green | strawberry level on glass |
| edge of the tray | drink is foamy | edge of tray |
| part of a shade | snack is lucious | part of shade |
| part of some leaves | background is dark | part of leaves |
| edge of another tray | cup is glass | |
| part of a leaf | shake is pink-colored | |
| The strawberries are lucious looking | | |



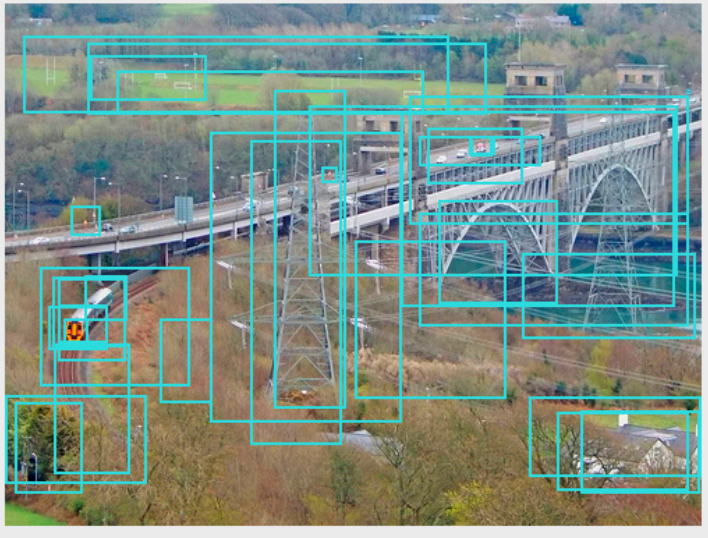
Question Answers

| | |
|---|---|
| Who is in the picture? | No one. |
| What kind of fruit is on the plate? | Strawberries. |
| When was the picture taken? | During the day. |
| What kind of drink is this? | Smoothie. |
| What fruits are shown? | Strawberries. |

Visual Genome: http://visualgenome.org/

# Question Answering

| Regions | Attributes | Relationships |
| --- | --- | --- |
| a train engine with a yellow front | grass is green | train has front |
| a train car being pulled by an engine | tracks is curving | engine pulling train |
| an electric tower holding up power lines | bush is brown | tower holds lines |
| a river crossing under a bridge | bush is dry | crossing under bridge |
| a bridge crossing over a river | structure is tall | crossing over river |
| a house near a river | structure is metal | house near river |
| a person in orange | water is blue | person on bridge |
| | | field behind bridge |
| | | field for soccer |
| | | tree near tracks |
| | | car on bridge |



Question Answers

| | |
| --- | --- |
| What is covering the view of the houses? | Trees. |
| Where is the river? | Under the bridge. |
| What color is the front of the train? | Orange. |
| How many arches on the bridge? | Two. |
| What is on most the trees? | Leaves. |

Visual Genome: http://visualgenome.org/

# Question Answering

```
1 Mary moved to the bathroom.
2 John went to the hallway.
3 Where is Mary?          bathroom          1
4 Daniel went back to the hallway.
5 Sandra moved to the garden.
6 Where is Daniel?        hallway           4
7 John moved to the office.
8 Sandra journeyed to the bathroom.
9 Where is Daniel?        hallway           4
10 Mary moved to the hallway.
11 Daniel travelled to the office.
12 Where is Daniel?       office            11
13 John went back to the garden.
14 John moved to the bedroom.
15 Where is Sandra?       bathroom          8
1 Sandra travelled to the office.
2 Sandra went to the bathroom.
3 Where is Sandra?        bathroom          2
```

Extract from the Facebook bAbI Dataset

# Human Question Answering

Imagine I gave you an article or an image,
asked you to memorize it, took it away, then
asked you various questions.
Even as intelligent as you are,
you're going to get a failing grade :(

Why?

- You can't store everything in working memory
- Without a question to direct your attention, you waste focus on unimportant details

**Optimal:** give you the input data, give you the question, allow as many glances as possible

# Think in terms of

# Information Bottlenecks

Where is your model forced to use a compressed representation?

Most importantly,
*is that a good thing*?

# Gated Recurrent Unit (GRU)

## Cho et al. 2014

$$h_t = GRU(x_t, h_{t-1})$$

- A type of recurrent neural network (RNN), similar to the LSTM
  - Consumes and/or generates sequences (chars, words, ...)
- The GRU updates an internal state *h* according to the:
  - **existing state *h*** and the **current input *x***

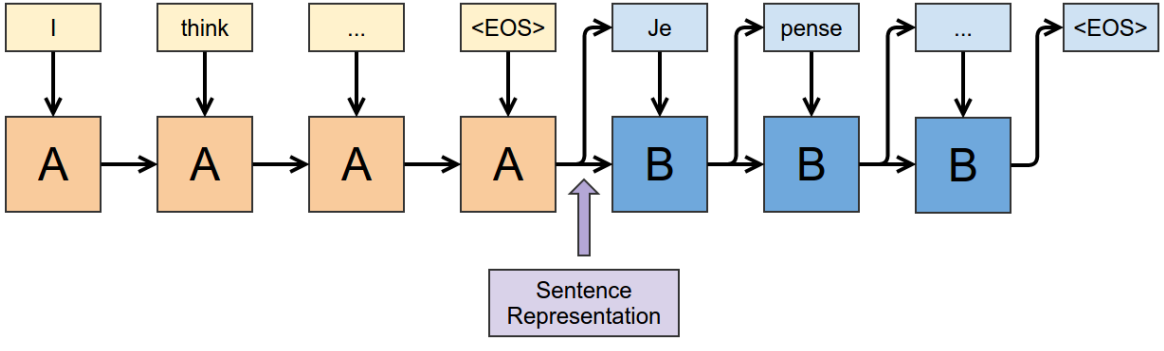



Figure from Chris Olah's Visualizing Representations

# Neural Machine Translation

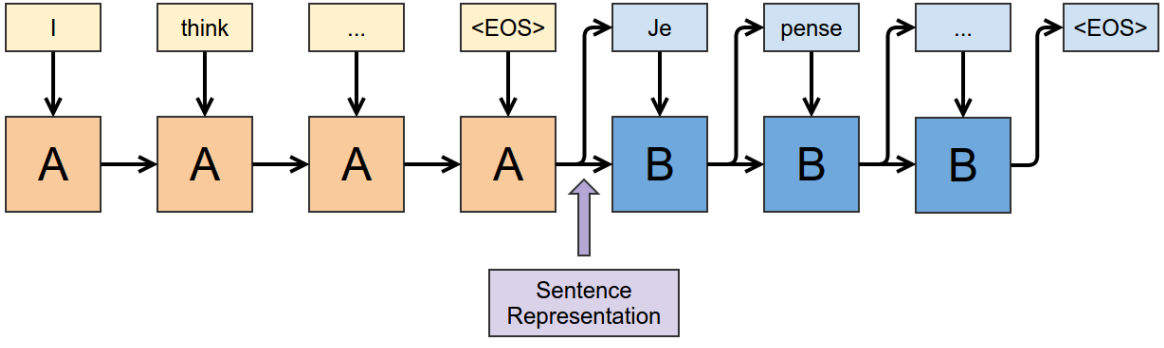



Figure from Chris Olah's Visualizing Representations

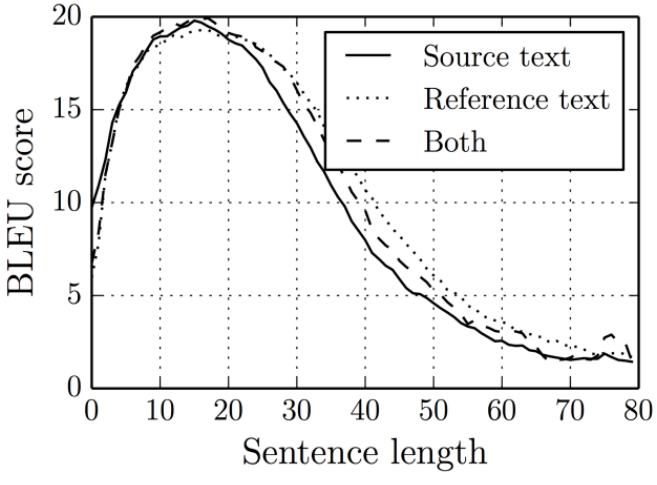
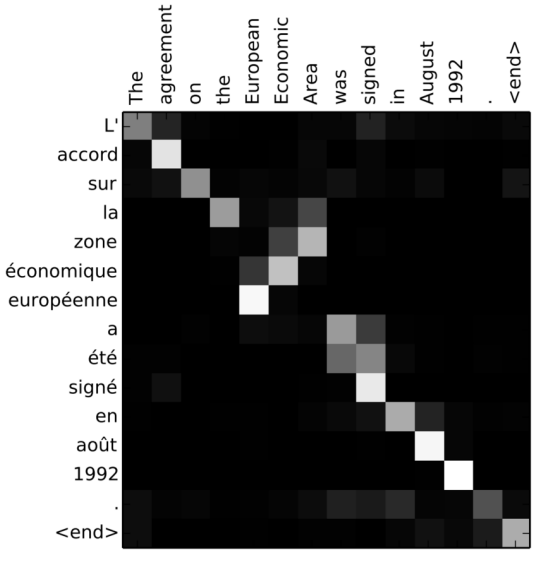
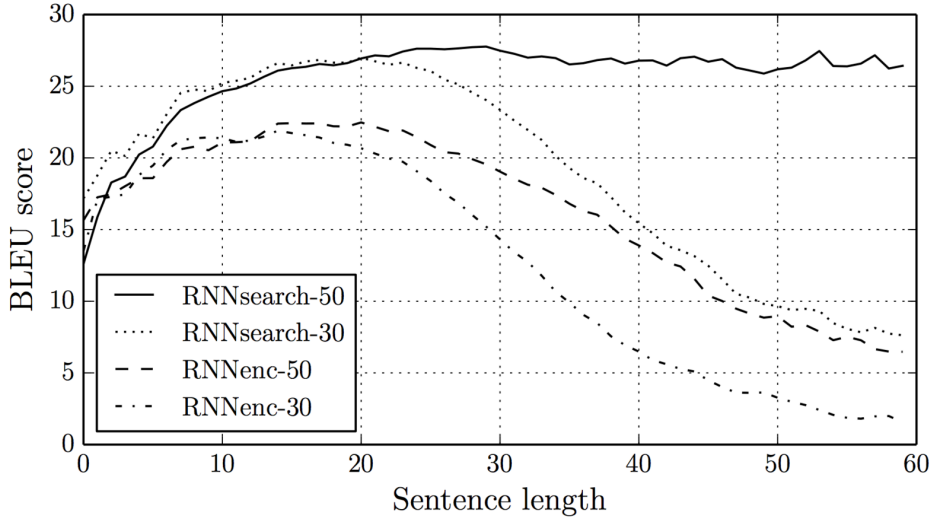



Figure from Bahdanau et al's
Neural Machine Translation by Jointly Learning to Align and Translate

# Neural Machine Translation




Results from Bahdanau et al's

Neural Machine Translation by Jointly Learning to Align and Translate

# Related Attention/Memory Work

- Sequence to Sequence (Sutskever et al. 2014)
- Neural Turing Machines (Graves et al. 2014)
- Teaching Machines to Read and Comprehend (Hermann et al. 2015)
- Learning to Transduce with Unbounded Memory (Grefenstette 2015)
- Structured Memory for Neural Turing Machines (Wei Zhang 2015)

- Memory Networks (Weston et al. 2015)
- End to end memory networks (Sukhbaatar et al. 2015)

# QA for Dynamic Memory Networks

- A modular and flexible DL framework for QA
- Capable of tackling wide range of tasks and input formats
- Can even been used for general NLP tasks (i.e. non QA) (PoS, NER, sentiment, translation, ...)

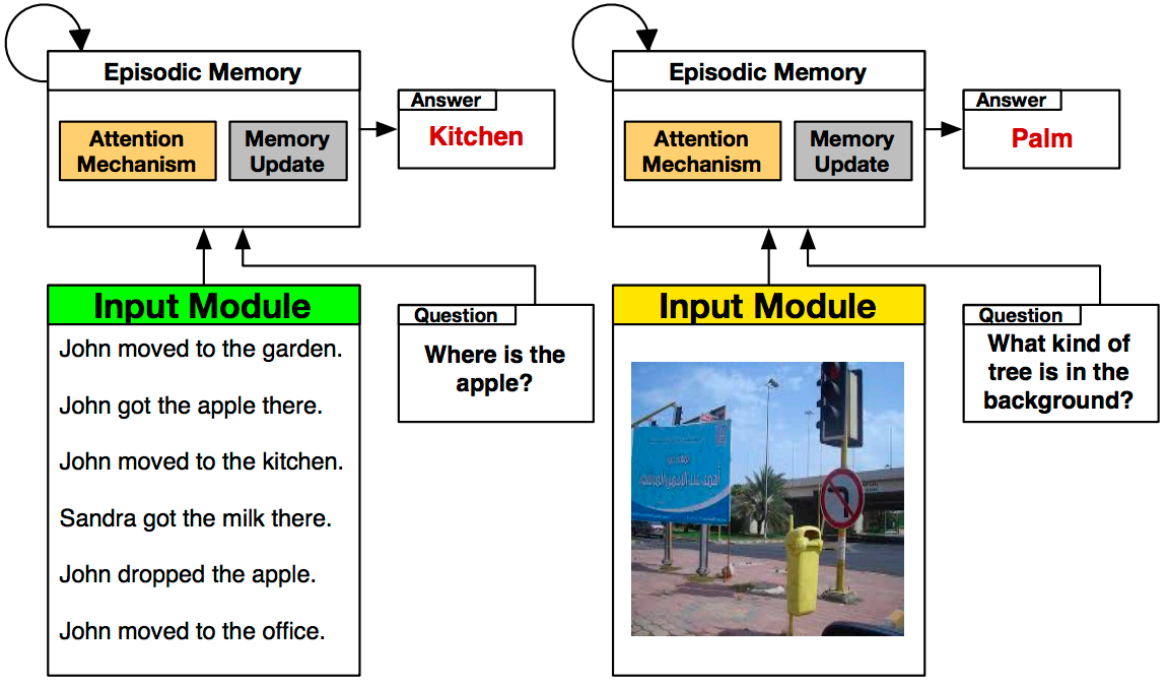
For full details:

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing (Kumar et al., 2015)

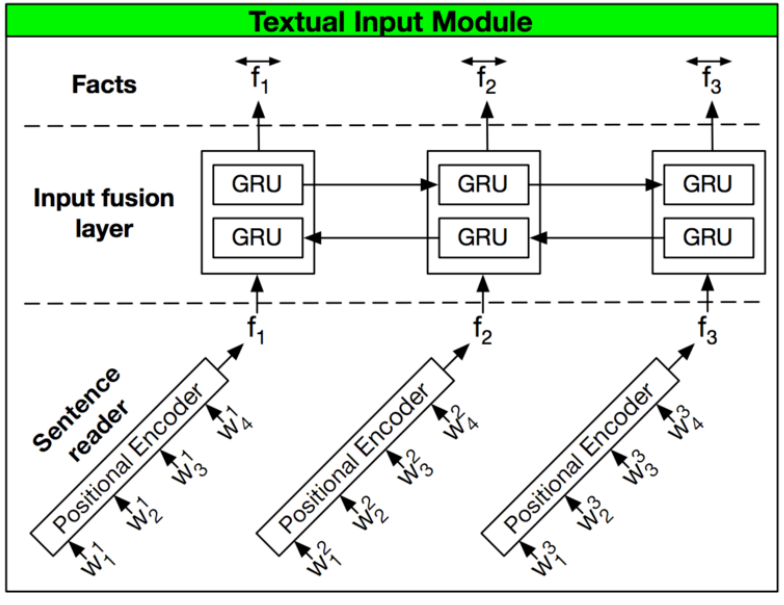
Dynamic Memory Networks for Visual and Textual Question Answering (Xiong et al., 2016)

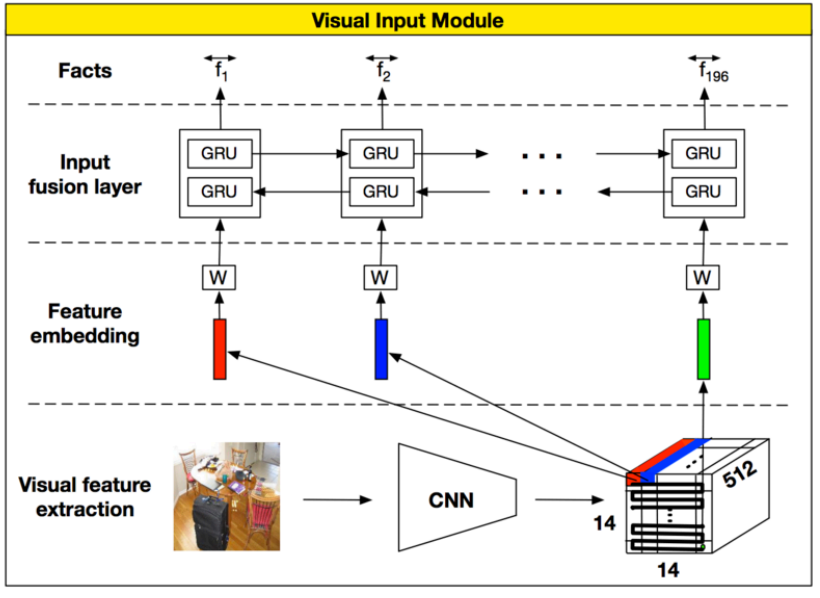
# QA for Dynamic Memory Networks

- A modular and flexible DL framework for QA
- Capable of tackling wide range of tasks and input formats
- Can even been used for general NLP tasks (i.e. non QA) (PoS, NER, sentiment, translation, ...)



# Input Modules



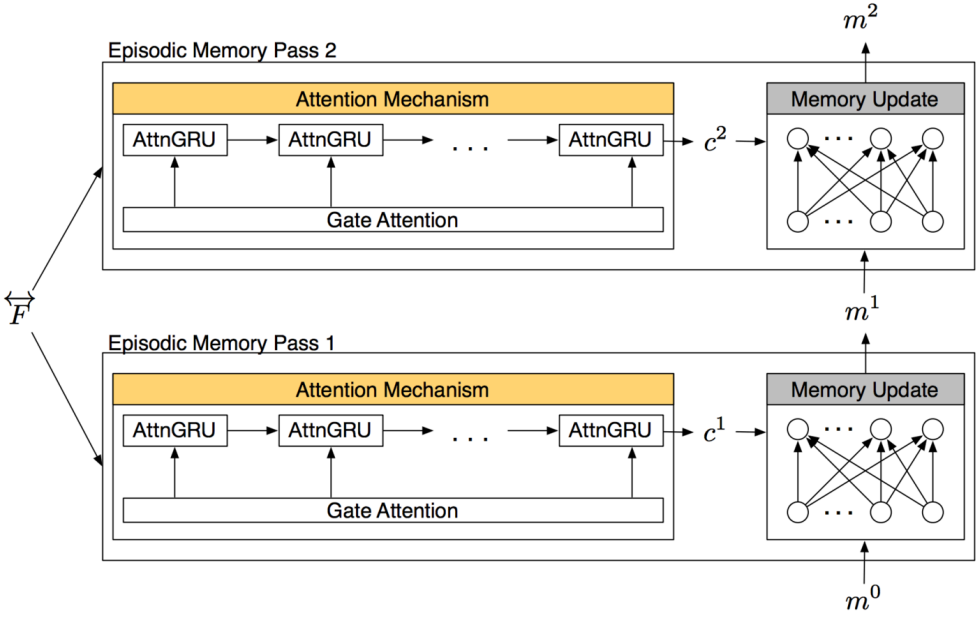




+ The module produces an ordered list of facts from the input
+ We can increase the number or dimensionality of these facts
+ Input fusion layer (bidirectional GRU) injects positional information and allows interactions between facts

# Episodic Memory Module

Composed of three parts with potentially multiple passes:

- Computing attention gates
- Attention mechanism
- Memory update



# Computing Attention Gates

Each fact receives an attention gate value from [0, 1]

The value is produced by analyzing [fact, query, episode memory]

Optionally enforce sparsity by using softmax over attention values

$$z_i^t = [\overleftrightarrow{f_i} \circ q; \overleftrightarrow{f_i} \circ m^{t-1}; |\overleftrightarrow{f_i} - q|; |\overleftrightarrow{f_i} - m^{t-1}|]$$

$$Z_i^t = W^{(2)} \tanh\left(W^{(1)} z_i^t + b^{(1)}\right) + b^{(2)}$$

$$g_i^t = \frac{\exp(Z_i^t)}{\sum_{k=1}^{M_i} \exp(Z_k^t)}$$

# Soft Attention Mechanism

Given the attention gates, we now want to extract a context vector from the input facts

$$c = \sum_{i=1}^{N} g_i f_i$$

If the gate values were passed through softmax,
the context vector is a weighted summation of the input facts

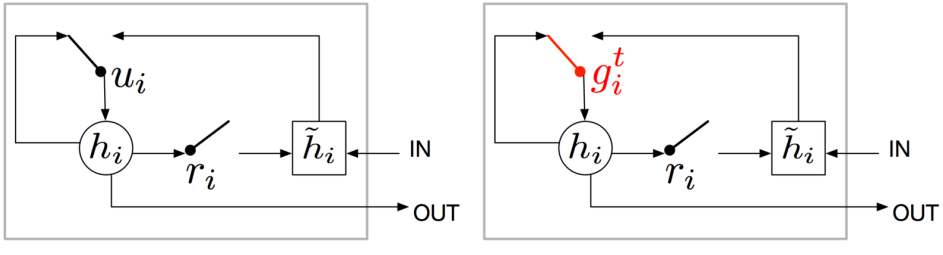
**Issue:** summation loses positional and ordering information

# Attention GRU Mechanism

If we modify the GRU, we can inject information from the attention gates.

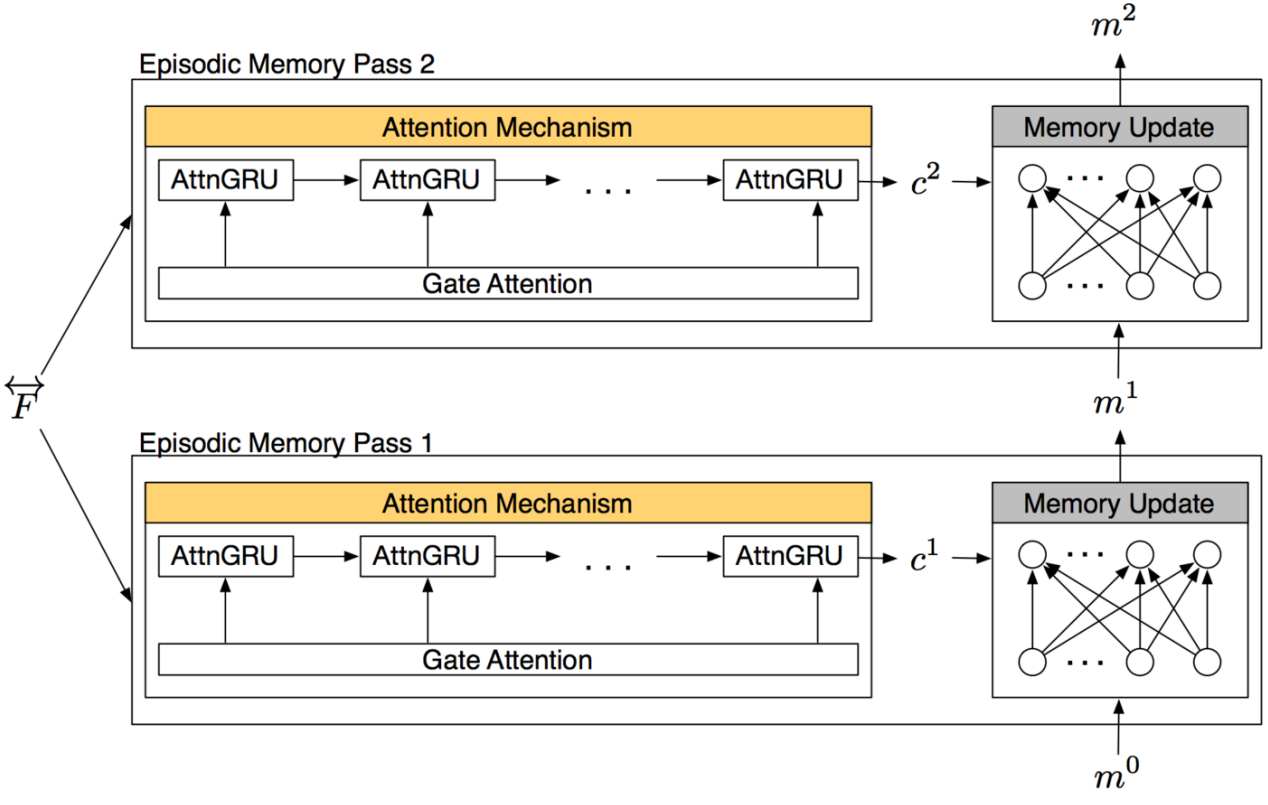
$$
\begin{aligned}
u_i &= \sigma\left(W^{(u)}x_i + U^{(u)}h_{i-1} + b^{(u)}\right) \\
r_i &= \sigma\left(W^{(r)}x_i + U^{(r)}h_{i-1} + b^{(r)}\right) \\
\tilde{h}_i &= \tanh\left(Wx_i + r_i \circ Uh_{i-1} + b^{(h)}\right) \\
h_i &= u_i \circ \tilde{h}_i + (1 - u_i) \circ h_{i-1}
\end{aligned}
$$

By replacing the update gate *u* with the activation gate *g*, the update gate can make use of the question and memory



# Attention GRU Mechanism

If we modify the GRU, we can inject information from the attention gates.



# For training,
# GPUs are leading the way

VisualQA dataset has over 200k images and 600k questions

- *GPUs are the key to efficient training, especially at higher resolutions*

The DMN make heavy use of RNNs

- *CNNs have experienced majority of optimization focus (many optimizations are trivial)*
- *RNNs on GPUs still have room to improve*
- *NVIDIA are actively improving RNN optimization*

# Results

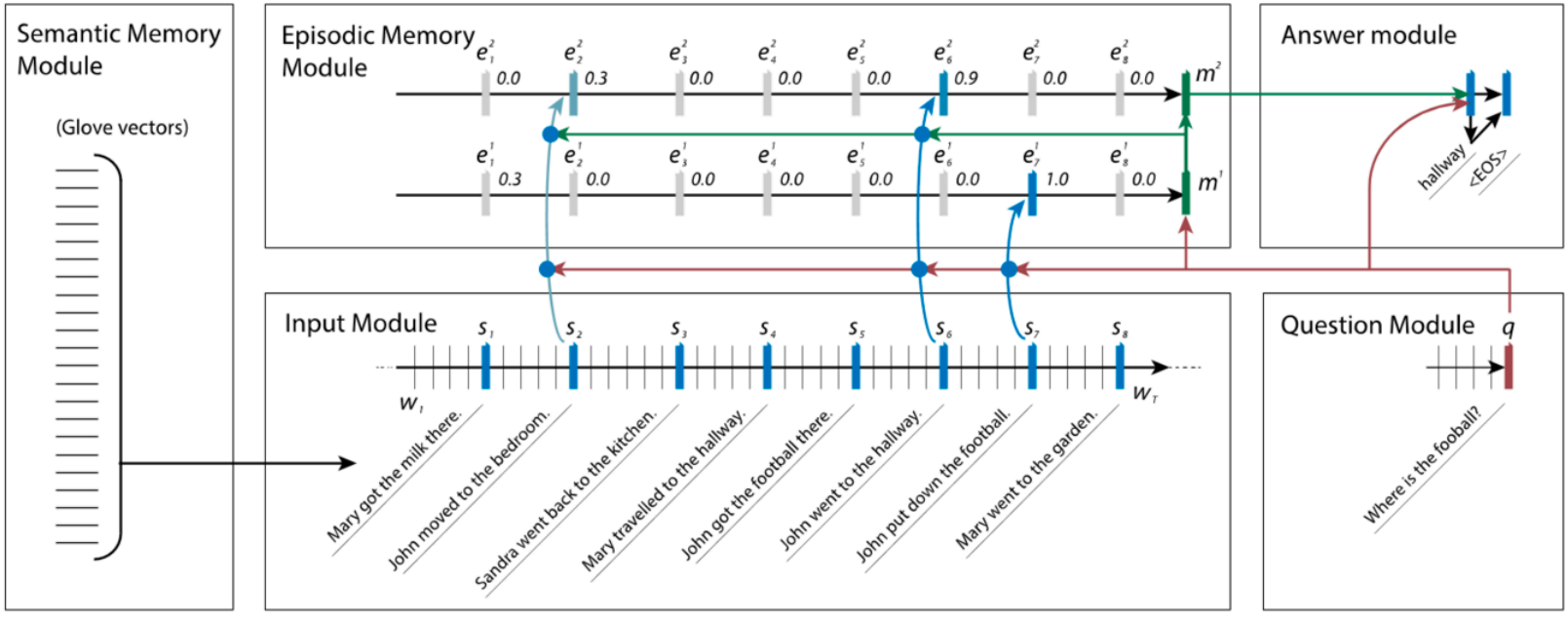
Focus on three experiments:

Text

Vision

Attention visualization

# DMN Overview



# Accuracy: Text QA (bAbI 10k)

| Task | DMN+ | E2E | NR |
|---|---|---|---|
| 2: 2 supporting facts | 0.3 | 0.3 | - |
| 3: 3 supporting facts | 1.1 | 2.1 | - |
| 5: 3 argument relations | 0.5 | 0.8 | - |
| 6: yes/no questions | 0.0 | 0.1 | - |
| 7: counting | 2.4 | 2.0 | - |
| 8: lists/sets | 0.0 | 0.9 | - |
| 9: simple negation | 0.0 | 0.3 | - |
| 11: basic coreference | 0.0 | 0.1 | - |
| 14: time reasoning | 0.2 | 0.1 | - |
| 16: basic induction | 45.3 | 51.8 | - |
| 17: positional reasoning | 4.2 | 18.6 | 0.9 |
| 18: size reasoning | 2.1 | 5.3 | - |
| 19: path finding | 0.0 | 2.3 | 1.6 |
| Mean error (%) | 2.8 | 4.2 | - |
| Failed tasks (err >5%) | 1 | 3 | - |

# Accuracy: Visual Question Answering

VQA test-dev and test-standard:

- Antol et al. (2015)
- ACK Wu et al. (2015);
- iBOWIMG - Zhou et al. (2015);
- DPPnet - Noh et al. (2015); D-NMN - Andreas et al. (2016);
- SAN - Yang et al. (2015)

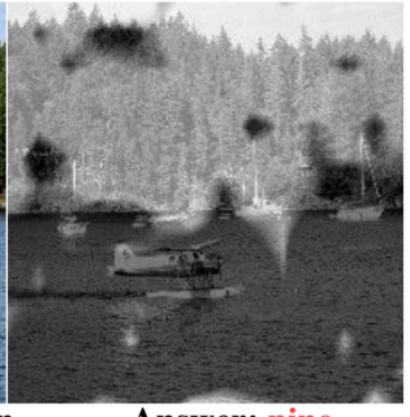
| | test-dev | | | | test-std |
|---|---|---|---|---|---|
| Method | All | Y/N | Other | Num | All |
| VQA | | | | | |
| Image | 28.1 | 64.0 | 3.8 | 0.4 | - |
| Question | 48.1 | 75.7 | 27.1 | 36.7 | - |
| Q+I | 52.6 | 75.6 | 37.4 | 33.7 | - |
| LSTM Q+I | 53.7 | 78.9 | 36.4 | 35.2 | 54.1 |
| ACK | 55.7 | 79.2 | 40.1 | 36.1 | 56.0 |
| iBOWIMG | 55.7 | 76.5 | 42.6 | 35.0 | 55.9 |
| DPPnet | 57.2 | 80.7 | 41.7 | 37.2 | 57.4 |
| D-NMN | 57.9 | 80.5 | 43.1 | 37.4 | 58.0 |
| SAN | 58.7 | 79.3 | 46.1 | 36.6 | 58.9 |
| DMN+ | **60.3** | 80.5 | 48.3 | 36.8 | **60.4** |

# Accuracy: Visual Question Answering

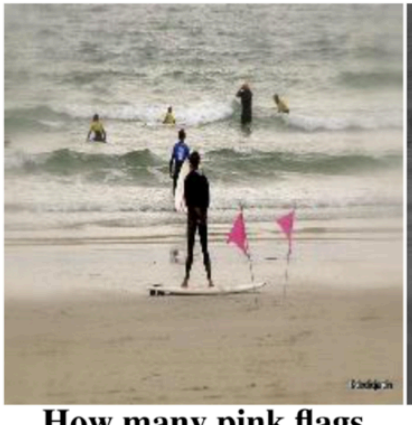
**What is the main color on the bus ?**

**Answer: blue**

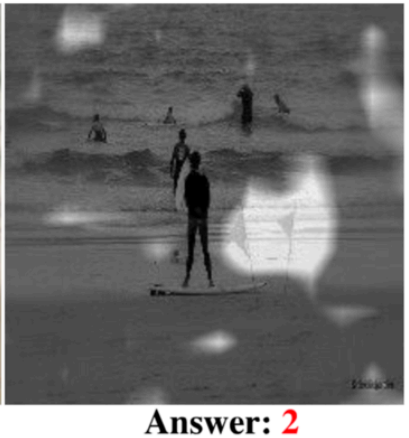
**What type of trees are in the background ?**



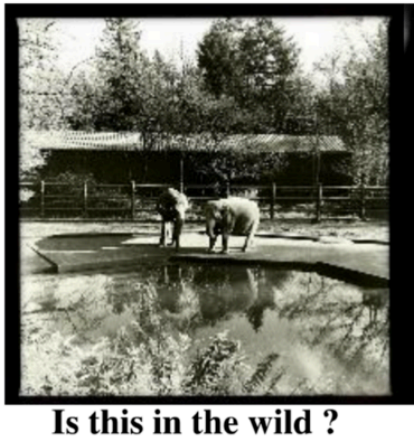
**Answer: pine**



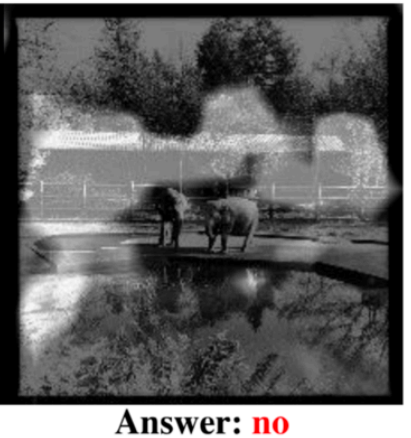
**How many pink flags are there ?**



**Answer: 2**



**Is this in the wild ?**



**Answer: no**

# Accuracy: Visual Question Answering

**Which man is dressed more flamboyantly ?** **Answer: right**


**Who is on both photos ?** **Answer: girl**

**What time of day was this picture taken ?** **Answer: night**

**What is the boy holding ?** **Answer: surfboard**

# Accuracy: Visual Question Answering

What is this sculpture made out of ?

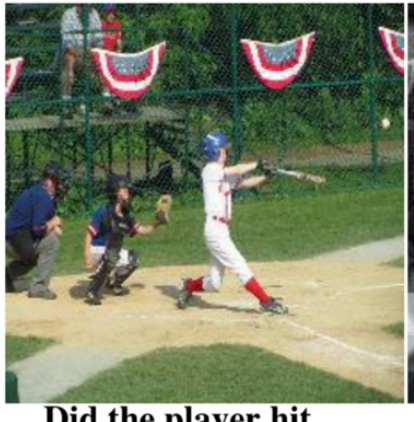
Answer: metal

What color are the bananas ?

Answer: green

What is the pattern on the cat ' s fur on its tail ?

Answer: stripes



Did the player hit the ball ?

Answer: yes

# Summary

- Attention and memory can avoid the information bottleneck
- The DMN can provide a flexible framework for QA work
- Attention visualization can help in model interpretability
- We have the compute power to explore all these!