Working Memory Networks:
Augmenting Memory Networks with a Relational Reasoning Module

Juan Pavez, Héctor Allende, Héctor Allende-Cid

UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

Reasoning for Question Answering

Reasoning is **crucial** for building systems that can dialogue with humans in natural language.

Reasoning: The process of forming conclusions, judgments, or inferences from facts or premises.

Examples:

- **Inferential Reasoning:** Premise 1, Premise 2 -> **Conclusion**
  - John is in the kitchen, John has the ball -> The ball is in the kitchen
- **Relational Reasoning:** Reason about relations between entities and their properties (Santoro et al.)
- **Causal Reasoning, Logical Reasoning, …**

bAbI Dataset (Weston et al., 2015)

- One of the earliest datasets to measure the reasoning abilities of ML systems.
- **Synthetic**. Not NLP.
- Easy to evaluate different **reasoning capabilities.**
- **Noiseless tasks: Separates** reasoning analysis from natural language understanding.
- A thorough analysis can be found in (Lee et al., 2016)

Category 2: Two Supporting Facts.
01: Mary went to the kitchen.
02: Sandra journeyed to the office
03: Mary got the football there.
04: Mary travelled to the garden.
05: Where is the football? garden 3 4

Category 4: Path Finding.
01: The bedroom is south of the hallway..
02: The bathroom is east of the office.
03: The kitchen is west of the garden.
04: The garden is south of the office.
05: The office is south of the bedroom.
05: How do you go from the garden to the bedroom?? n,n 4 5

bAbI Dataset (Weston et al., 2015)

- One of the earliest datasets to measure the reasoning abilities of ML systems.
- **Synthetic**. Not NLP.
- Easy to evaluate different **reasoning capabilities.**
- **Noiseless tasks: Separates** reasoning analysis from natural language understanding.
- A thorough analysis can be found in (Lee et al., 2016)

Category 2: Two Supporting Facts.
01: Mary went to the kitchen.
02: Sandra journeyed to the office
03: Mary got the football there.
04: Mary travelled to the garden.
05: Where is the football? garden 3 4

$$Has(Mary, Football), \; Is(Mary, Garden) \rightarrow Is(Football, Garden)$$

Category 4: Path Finding.
01: The bedroom is south of the hallway..
02: The bathroom is east of the office.
03: The kitchen is west of the garden.
04: The garden is south of the office.
05: The office is south of the bedroom.
05: How do you go from the garden to the bedroom?? n,n 4 5

bAbI Dataset (Weston et al., 2015)

- One of the earliest datasets to measure the reasoning abilities of ML systems.
- **Synthetic**. Not NLP.
- Easy to evaluate different **reasoning capabilities.**
- **Noiseless tasks: Separates** reasoning analysis from natural language understanding.
- A thorough analysis can be found in (Lee et al., 2016)

Category 2: Two Supporting Facts.
01: Mary went to the kitchen.
02: Sandra journeyed to the office
03: Mary got the football there.
04: Mary travelled to the garden.
05: Where is the football? garden 3 4

Category 4: Path Finding.
01: The bedroom is south of the hallway..
02: The bathroom is east of the office.
03: The kitchen is west of the garden.
04: The garden is south of the office.
05: The office is south of the bedroom.
05: How do you go from the garden to the bedroom?? n,n 4 5

$$S(Garden, Office), S(Office, Bedroom)$$
$$\& \ N = S^{-1}$$
$$\rightarrow N, N$$

Memory Augmented Neural Networks

Memory Networks (Weston et al. 2014, Sukhbaatar et al. 2015)

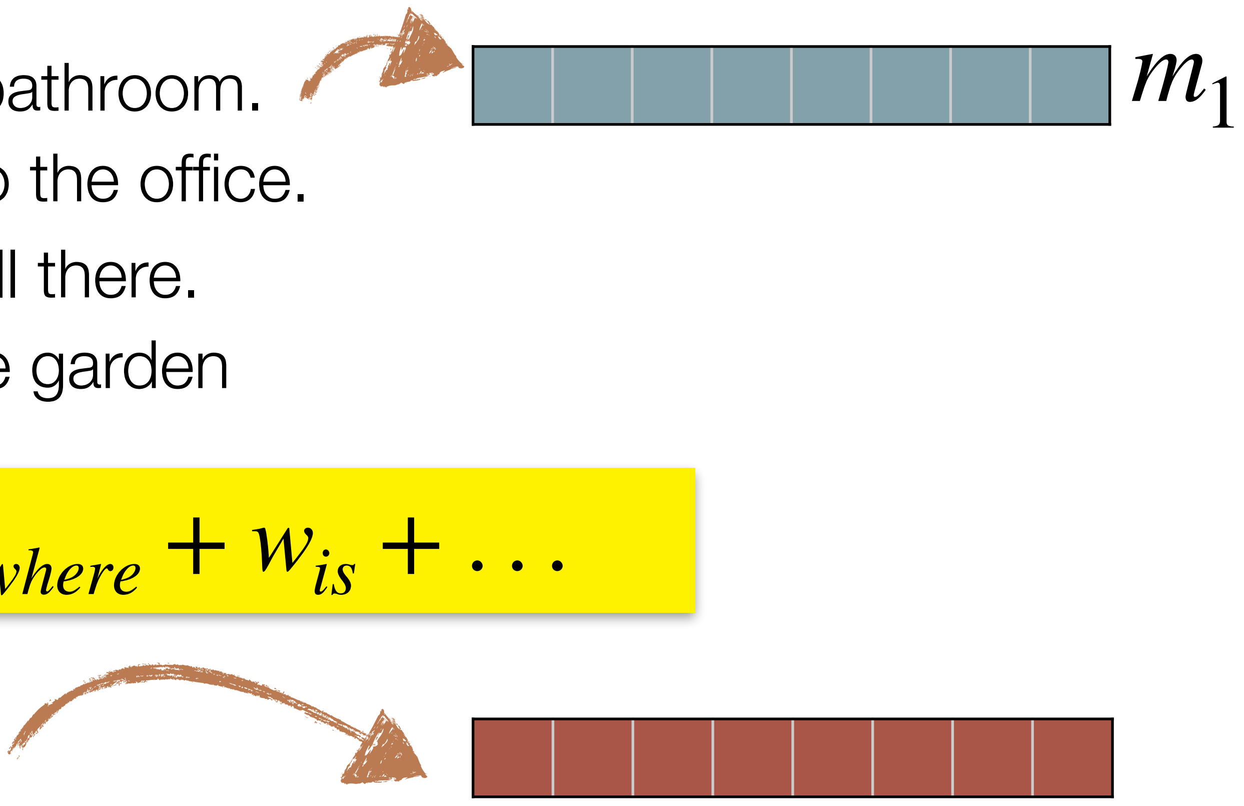
Process a set of inputs and store them in memory. Then, at each hop, an important part of the memory is retrieved and used to retrieve more memories. Finally, the last retrieved memory is used to compute the answer.

$$m_1 = w_{daniel} + w_{went} + \ldots$$

01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden

$$u = w_{where} + w_{is} + \ldots$$

q:Where is the football.



Memory Augmented Neural Networks

Memory Networks (Weston et al. 2014, Sukhbaatar et al. 2015)

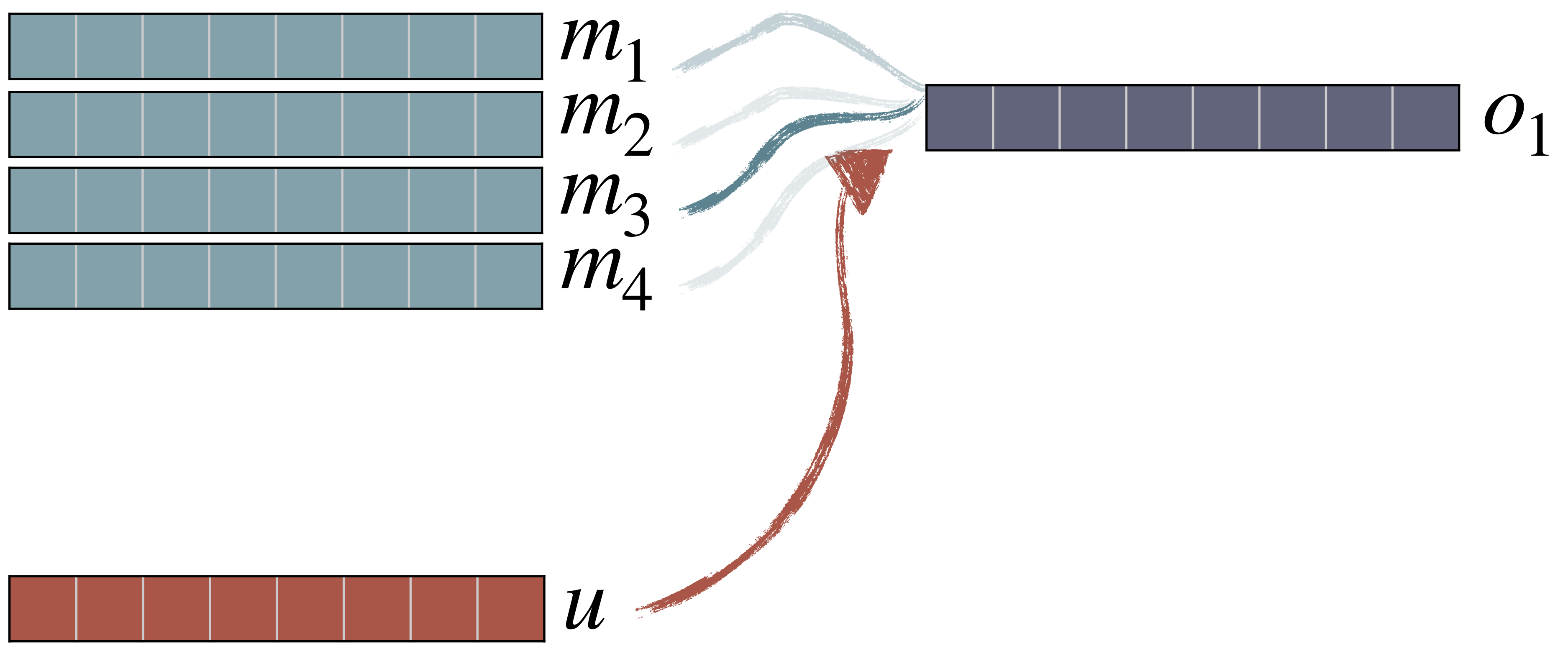
Process a set of inputs and store them in memory. Then, at each hop, an important part of the memory is retrieved and used to retrieve more memories. Finally, the last retrieved memory is used to compute the answer.

$$\alpha_i = Softmax(u^T m_i)$$

$$o_1 = \sum_i \alpha_i m_i$$

01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden

q:Where is the football.



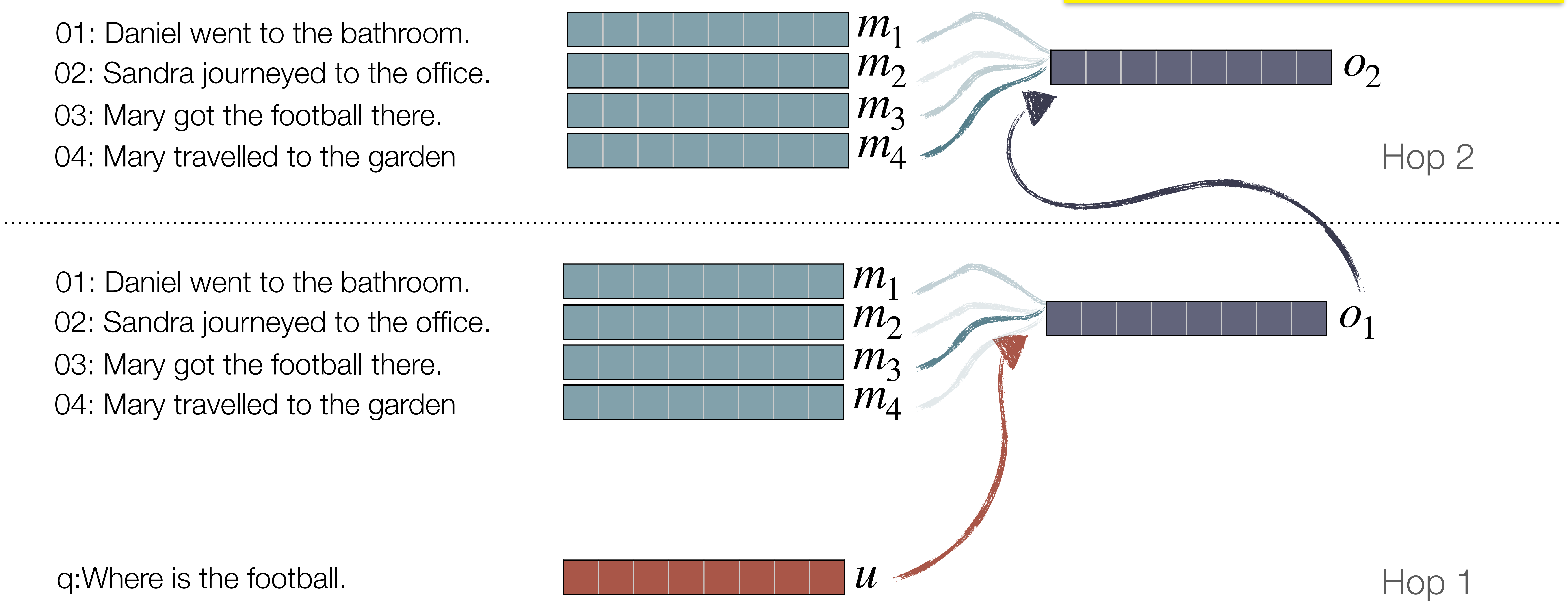
Memory Augmented Neural Networks

Memory Networks (Weston et al. 2014, Sukhbaatar et al. 2015)

Process a set of inputs and store them in memor[...] important part of the memory is retrieved and used to retrieve more m[...]retrieved memory is used to compute the answer.

$$\alpha_i = Softmax(o_1^T m_i)$$

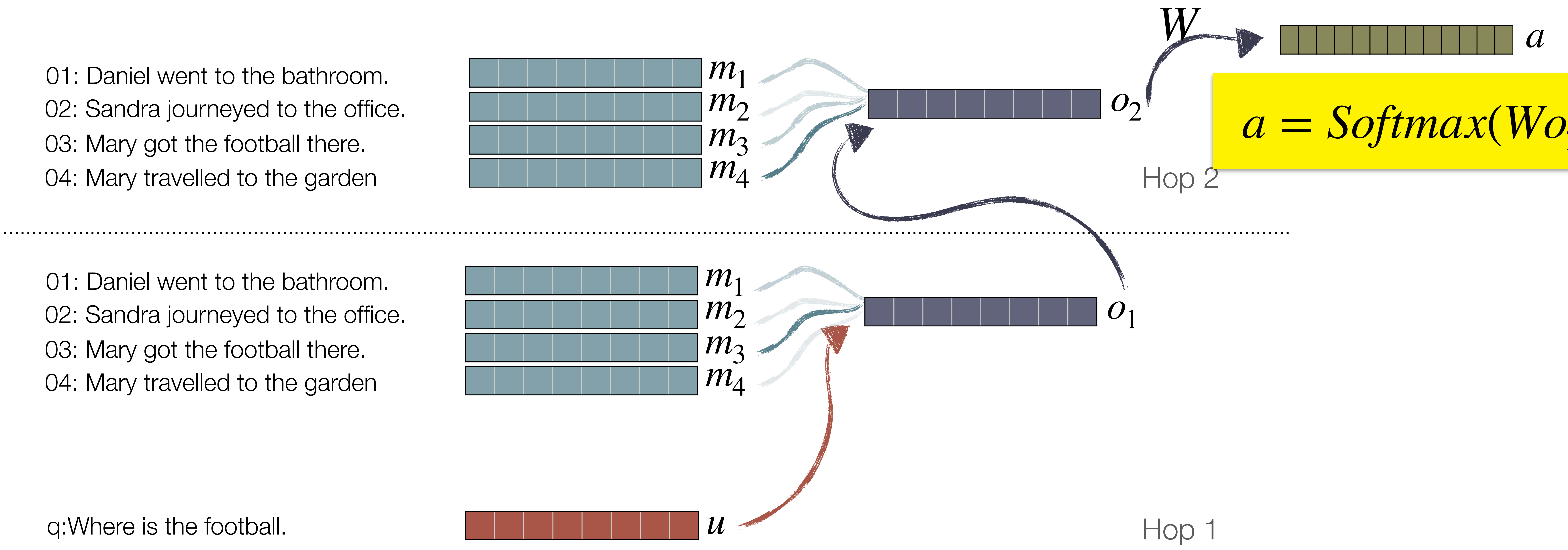
$$o_2 = \sum_i \alpha_i m_i$$



Memory Augmented Neural Networks

Memory Networks (Weston et al. 2014, Sukhbaatar et al. 2015)

Process a set of inputs and store them in memory. Then, at each hop, an important part of the memory is retrieved and used to retrieve more memories. Finally, the last retrieved memory is used to compute the answer.



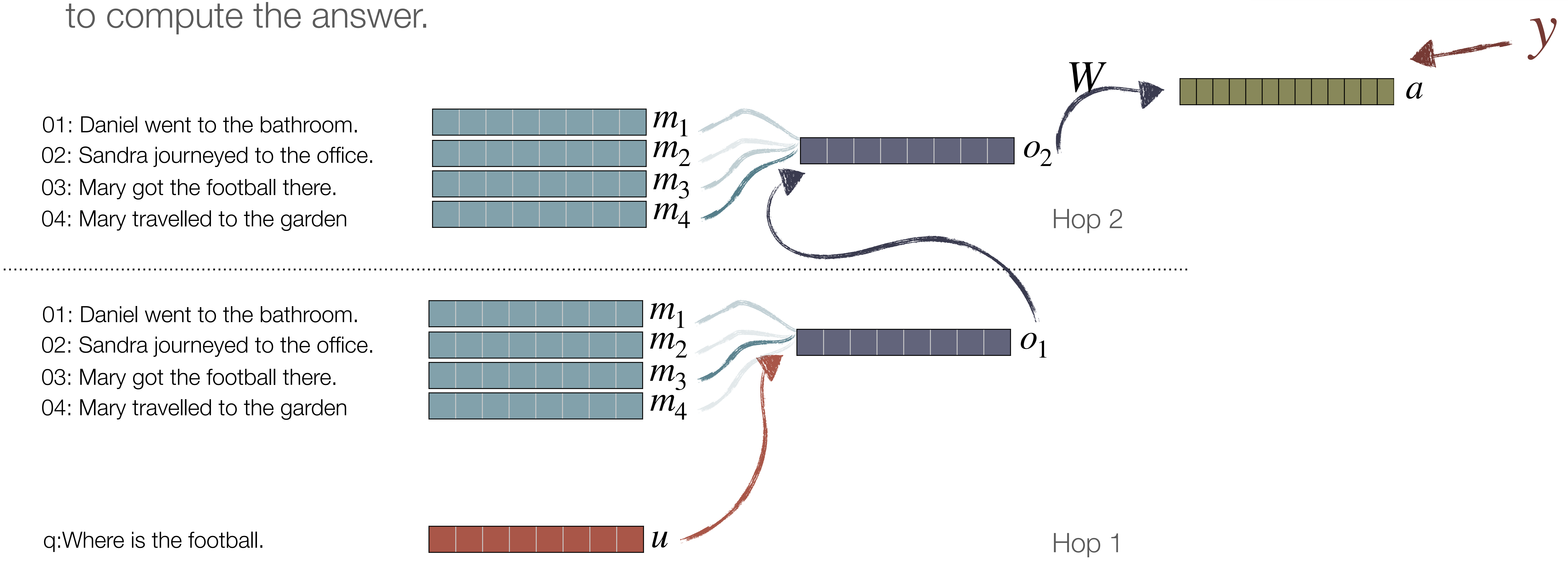
$$a = Softmax(Wo_2)$$

Memory Augmented Neural Networks

Memory Networks (Weston et al. 2014, Sukhbaatar et al. 20

Process a set of inputs and store them in memory. Then, at each hop, an imp memory is retrieved and used to retrieve more memories. Finally, the last retrie to compute the answer.

$$L(y, \hat{y}) = -\sum_i y_i ln(\hat{y}_i)$$



Memory Augmented Neural Networks

Memory Networks (Weston et al. 2014, Sukhbaatar et al. 2015)

Some weaknesses:

- The attention mechanism is **simple**
- The attention mechanism **relies on embeddings**.
  - It may be nice to separate embedding learning from attention learning (**modularization, reusability**).
- The **answer computation is too simple**, it only uses one retrieved memory. Hard to see how can produce more complex reasoning based on memories.

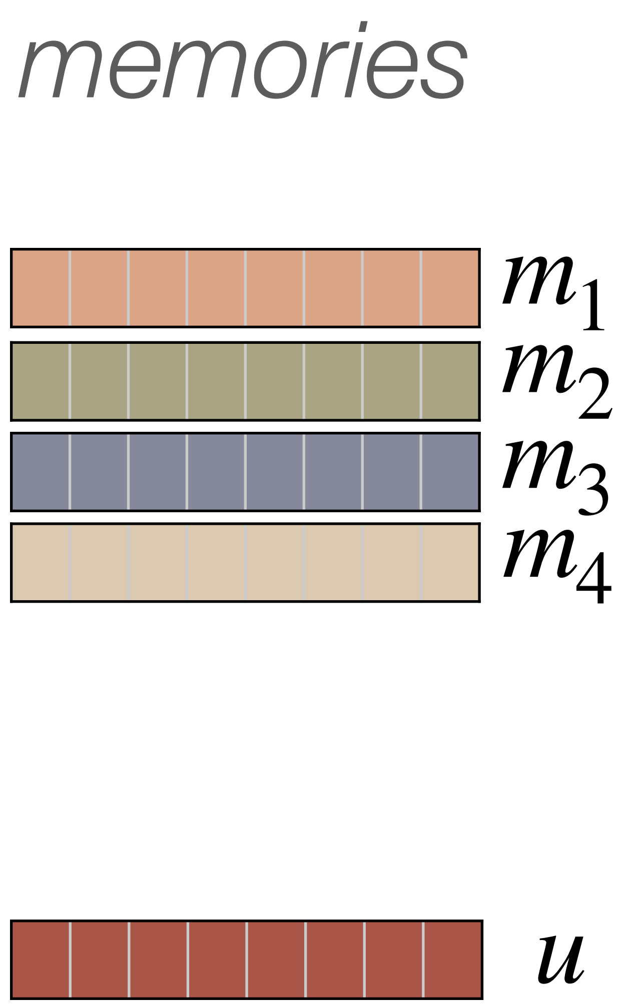
Relational Neural Networks

Relation Networks (Santoro et al. 2017)

Neural Network with an inductive bias to learn pairwise relations of the input objects and their properties. A type of Graph Neural Networks.

01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden

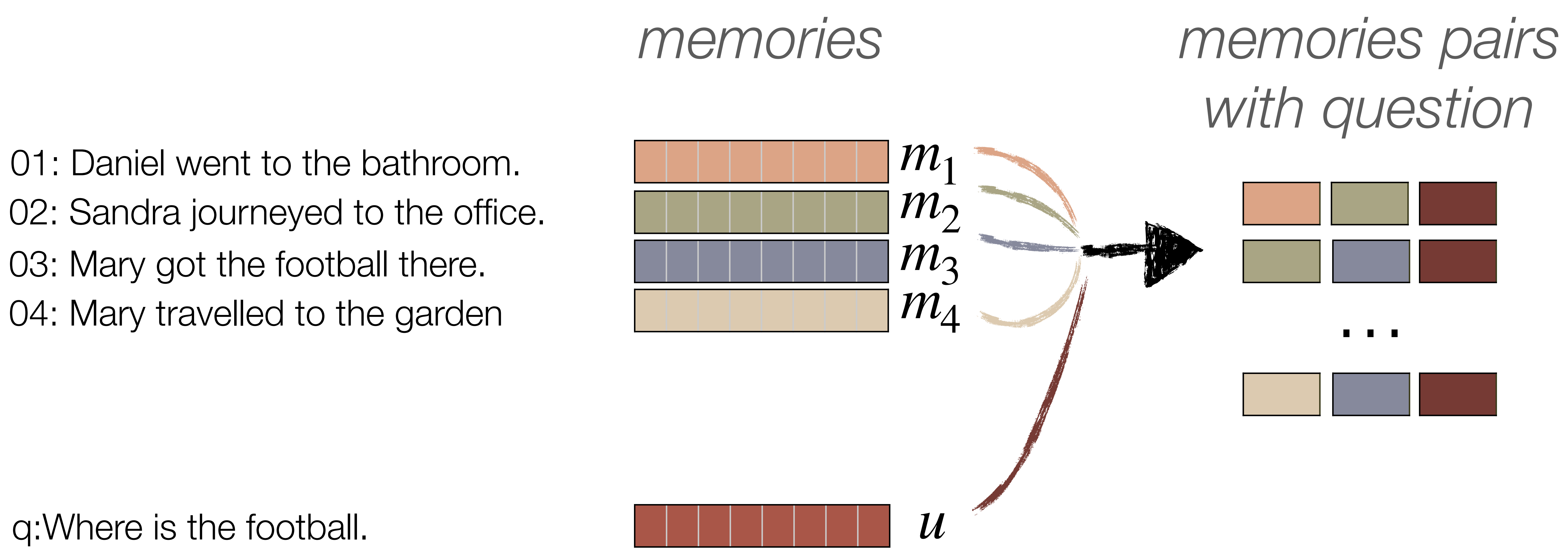
q:Where is the football.



Relational Neural Networks

Relation Networks (Santoro et al. 2017)

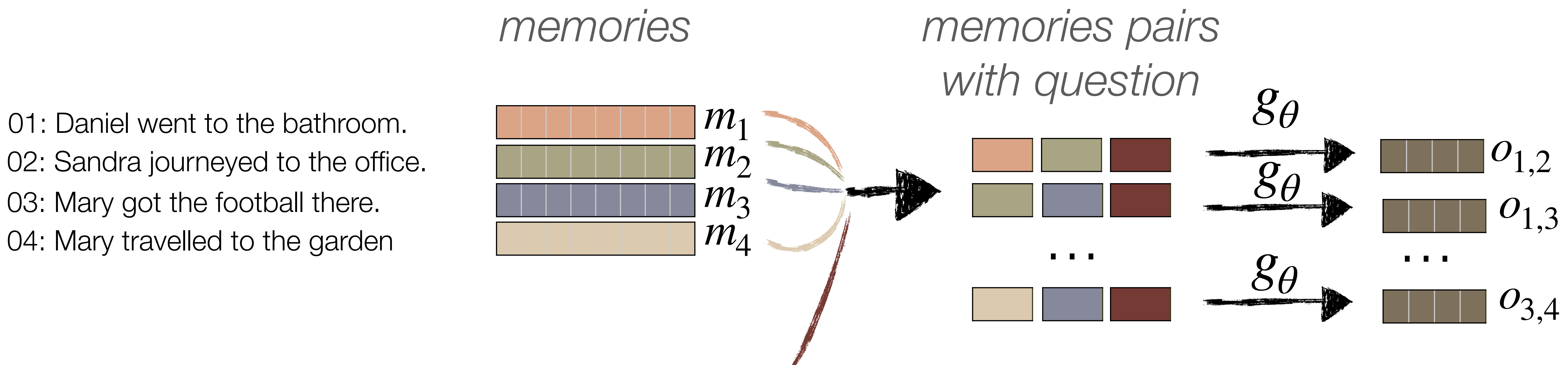
Neural Network with an inductive bias to learn pairwise relations of the input objects and their properties. A type of Graph Neural Networks.



Relational Neural Networks

Relation Networks (Santoro et al. 2017)

Neural Network with an inductive bias to learn pairwise relations of the input objects and their properties. A type of Graph Neural Networks.

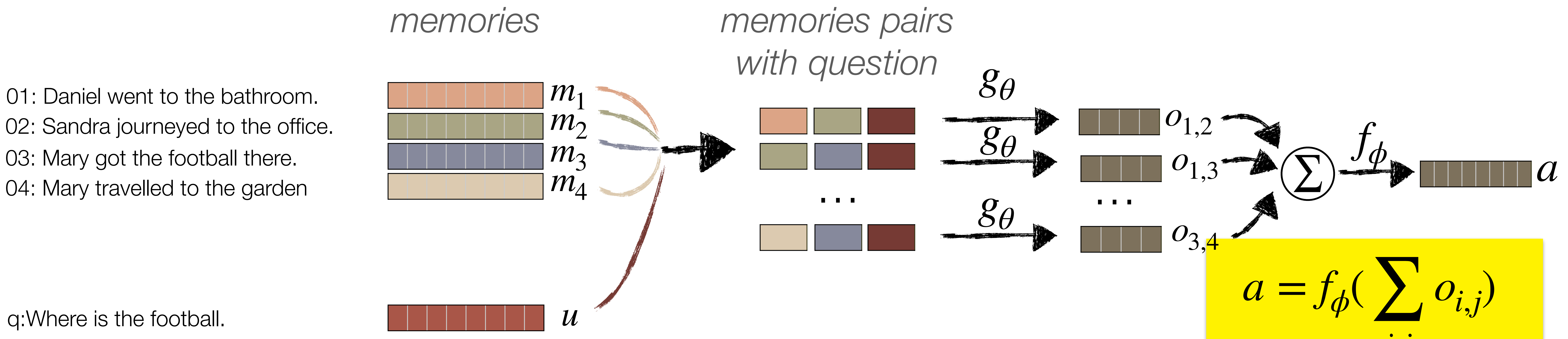


$$o_{i,j} = g_\theta([m_i; m_j; u])$$

Relational Neural Networks

Relation Networks (Santoro et al. 2017)

Neural Network with an inductive bias to learn pairwise relations of the input objects and their properties. A type of Graph Neural Networks.



$$a = f_\phi(\sum_{i,j} o_{i,j})$$

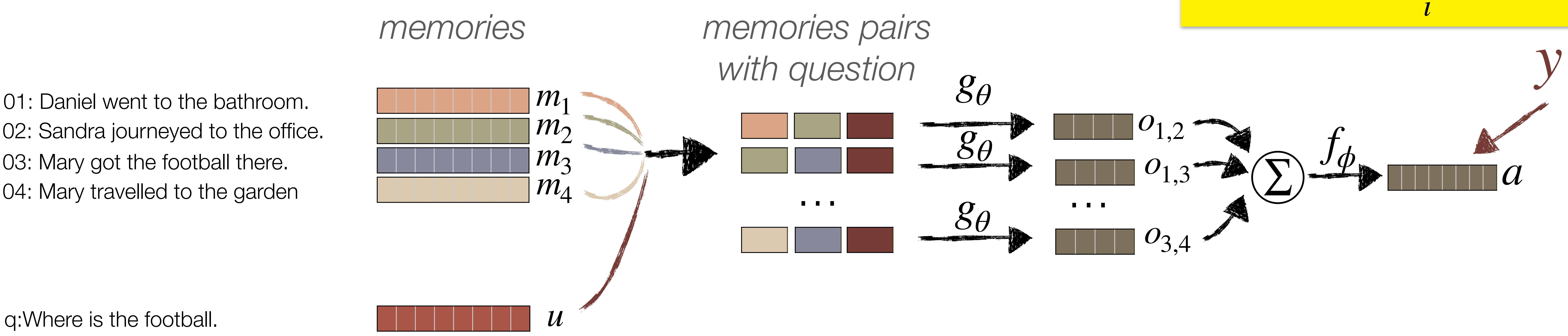
Relational Neural Networks

Relation Networks (Santoro et al. 2017)

Neural Network with an inductive bias to learn pairwise relations of the input ob

properties. A type of Graph Neural Networks.

$$L(y, \hat{y}) = -\sum_i y_i ln(\hat{y}_i)$$



Relational Neural Networks

Relation Networks (Santoro et al. 2017)

Some weaknesses:

- The model needs to process N^2 pairs where N is the number of memories.
  - 500 memories would require 250k backward and forward computations!
- Can not filter out unuseful objects that can produce **spurious relations**.

Working Memory Networks

Working Memory Network (Pavez et al., 2018)

A Memory Network model with a new working memory buffer and relational reasoning module. Produces state-of-the-art results in reasoning tasks. Inspired by the Multi-component model of working memory.

Short-term Memory Module

$$m_1 = GRU([W_{daniel}; W_{went}; \ldots])$$

01: Daniel went to the bathroom. m_1

02: Sandra journeyed to the office.

03: Mary got the football there.

04: Mary travelled to the garden

$$u = GRU([W_{where}; W_{is}; \ldots])$$

q:Where is the football. u

Attention Module

Reasoning Module

Working Memory Networks

Working Memory Network (Pavez et al., 2018)

A Memory Network model with a new working memory buffer and relational reasoning module. Produces state-of-the-art results in reasoning tasks. Inspired by the Multi-component model of working memory.

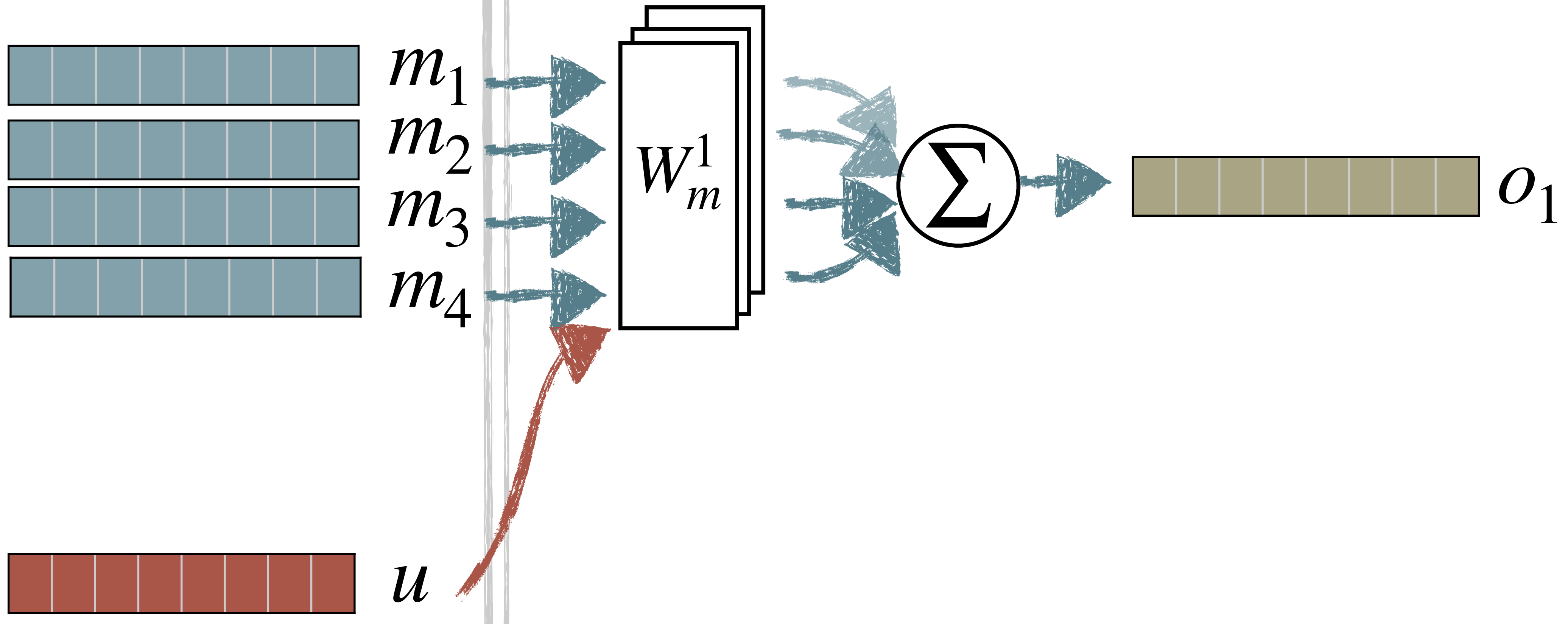
Short-term Memory Module

01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden

q:Where is the football.

Multi-head attention (Vaswani et al. 2017)

$$m_i^l = W_m^l m_i$$
$$\alpha_i^l = Softmax((u^T m_i^l / \sqrt{d}))$$
$$h_l = \sum_j \alpha_j^l m_j^l$$
$$o_1 = [h_1; h_2; \ldots] W_o$$



Reasoning Module

Working Memory Networks

Working Memory Netwo

A Memory Network model with a d relational reasoning module. Produces state-of-the-art results in reasonin omponent model of working memory.

Multi-head attention (Vaswani et al. 2017)

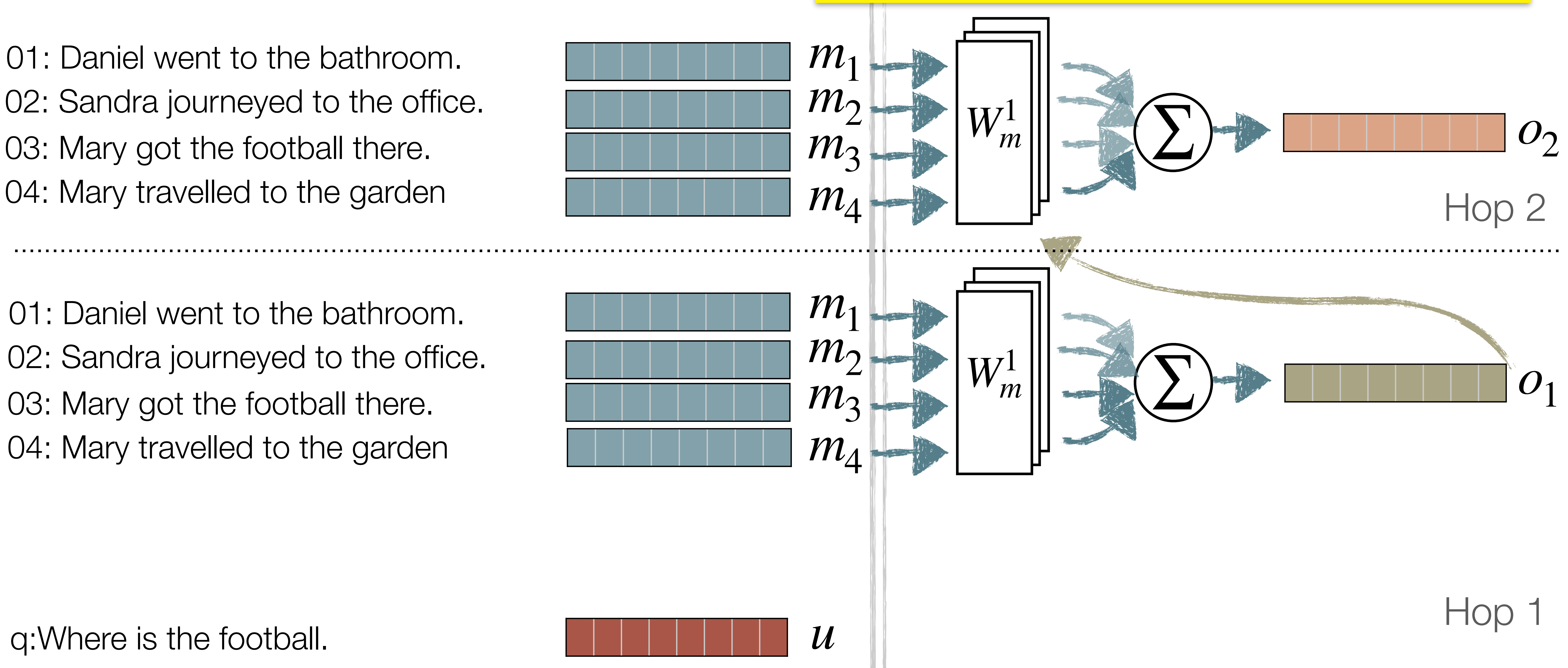
$$m_i^l = W_m^l m_i$$

$$\alpha_i^l = Softmax(((f_t(o_1)^T m_i^l / \sqrt{d}))$$

$$h_l = \sum_j \alpha_j^l m_j^l$$

$$o_1 = [h_1; h_2; \ldots]W_o$$

Short-term Memory Module



Reasoning Module

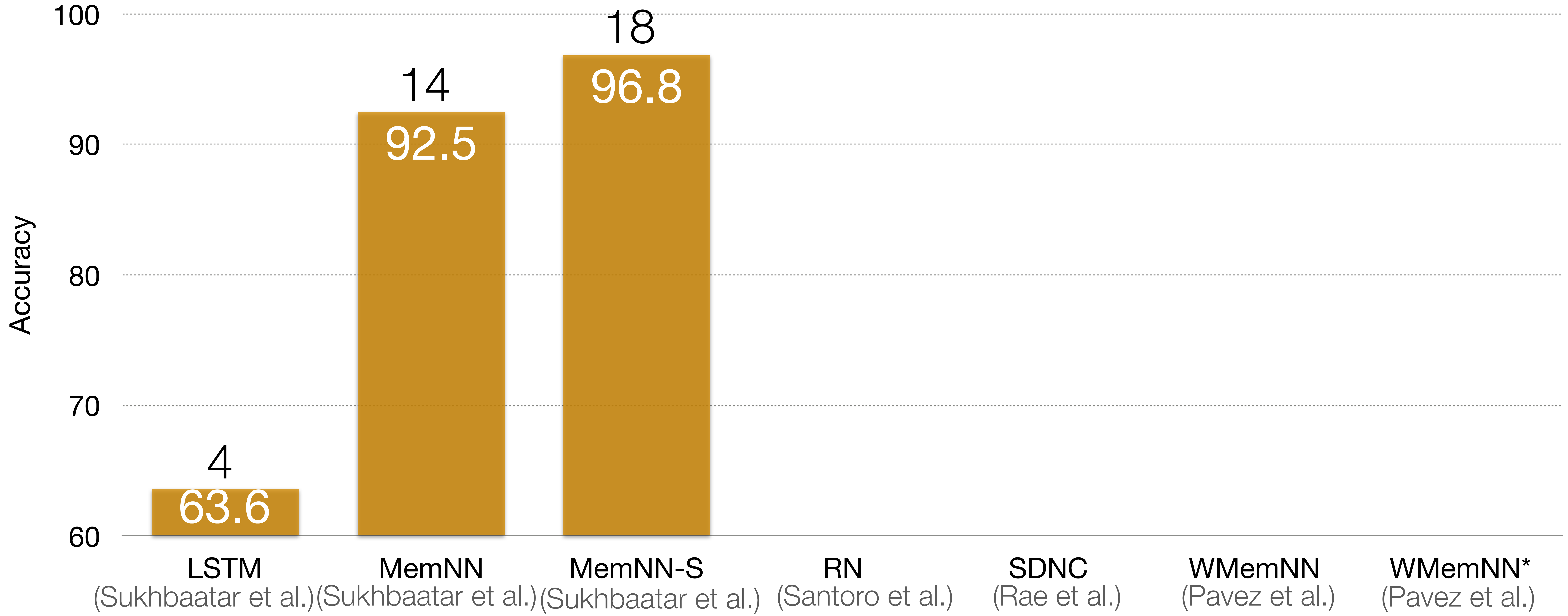

Short-term Memory Module
Attention Module
Reasoning Module
01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden
m_1
m_2
m_3
m_4
W_m^1
Σ
Working memory buffer
o_H
Hop H
...
f_t
01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden
m_1
m_2
m_3
m_4
W_m^1
Σ
o_2
Hop 2
01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden
m_1
m_2
m_3
m_4
W_m^1
Σ
o_1
Hop 1
q:Where is the football.
u


Short-term Memory Module
Attention Module
Reasoning Module
01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden
m_1 m_2 m_3 m_4
W_m^1
Working memory buffer
o_H
Hop H
...
memories pairs with question
o_2
Hop 2
g_θ
$o_{4,H}$
$o_{3,4}$
$o_{1,2}$
f_ϕ
a
01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden
o_1
$a = f_\phi(\sum_{i,j} g_\theta([o_i; o_j; u]))$
q:Where is the football.
u
Hop 1


Short-term Memory Module
Attention Module
Reasoning Module
01: Daniel went to the bathroom.
02: Sandra journeyed to the office.
03: Mary got the football there.
04: Mary travelled to the garden
m_1
m_2
m_3
m_4
W_m^1
$\sum$
Working memory buffer
o_H
Hop H
...
o_2
Hop 2
o_1
Hop 1
q:Where is the football.
u
$L(y, \hat{y}) = -\sum_i y_i ln(\hat{y}_i)$
memories pairs with question
g_θ
$o_{4,H}$
$o_{3,4}$
$o_{1,2}$
f_ϕ
y
a

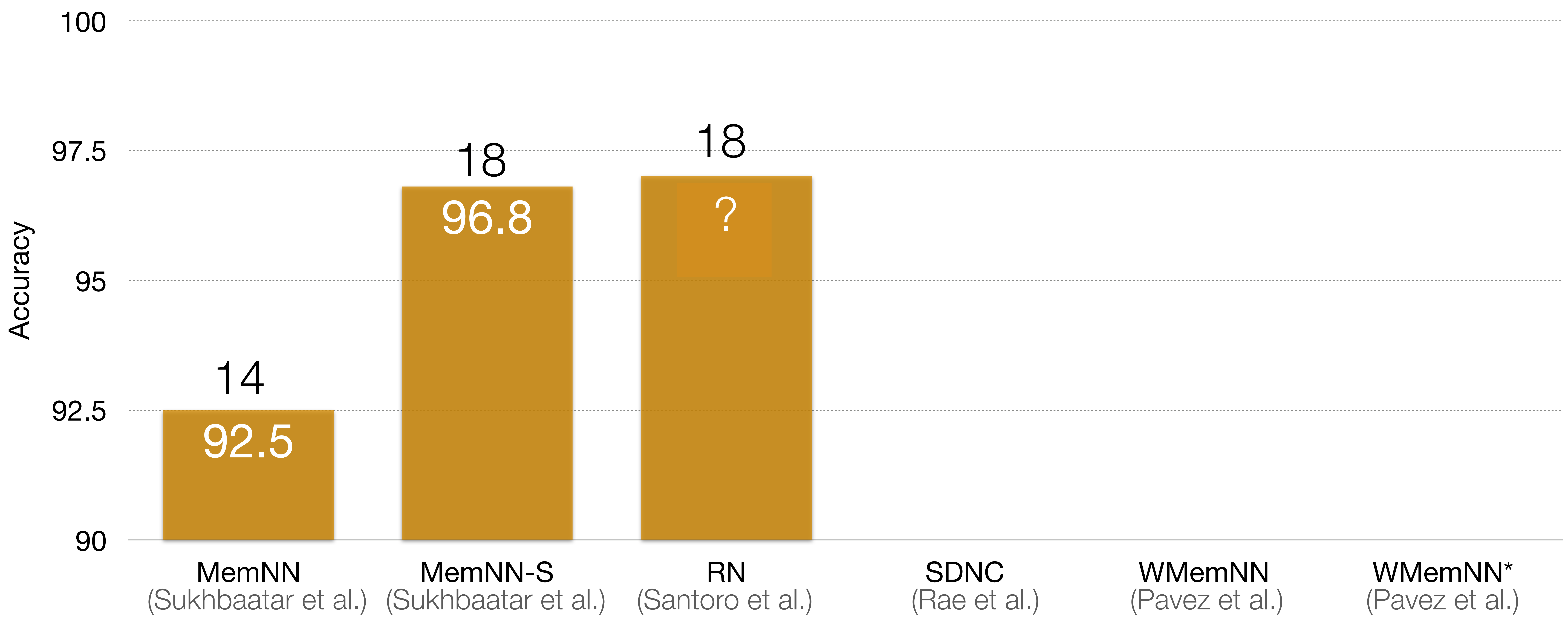
Results - Jointly trained bAbI-10k.

- Results on **jointly trained** bAbI-10k: Train a single model on all tasks simultaneously.
- Note that **EntNet** (Henaff et al.) solves all tasks in the **per-task version**: A single model for each task.



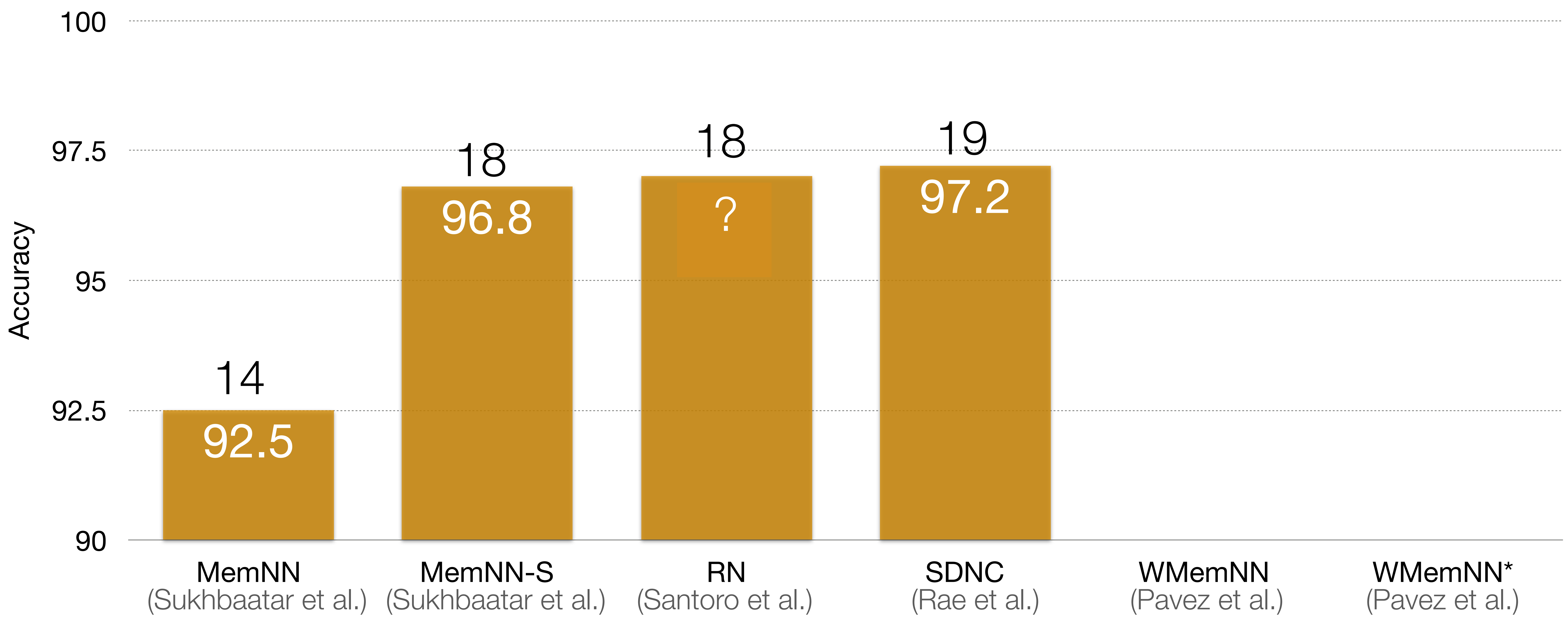
Results - Jointly trained bAbI-10k.

- Results on **jointly trained** bAbI-10k: Train a single model on all tasks simultaneously.
- Note that **EntNet** (Henaff et al.) solves all tasks in the **per-task version**: A single model for each task.



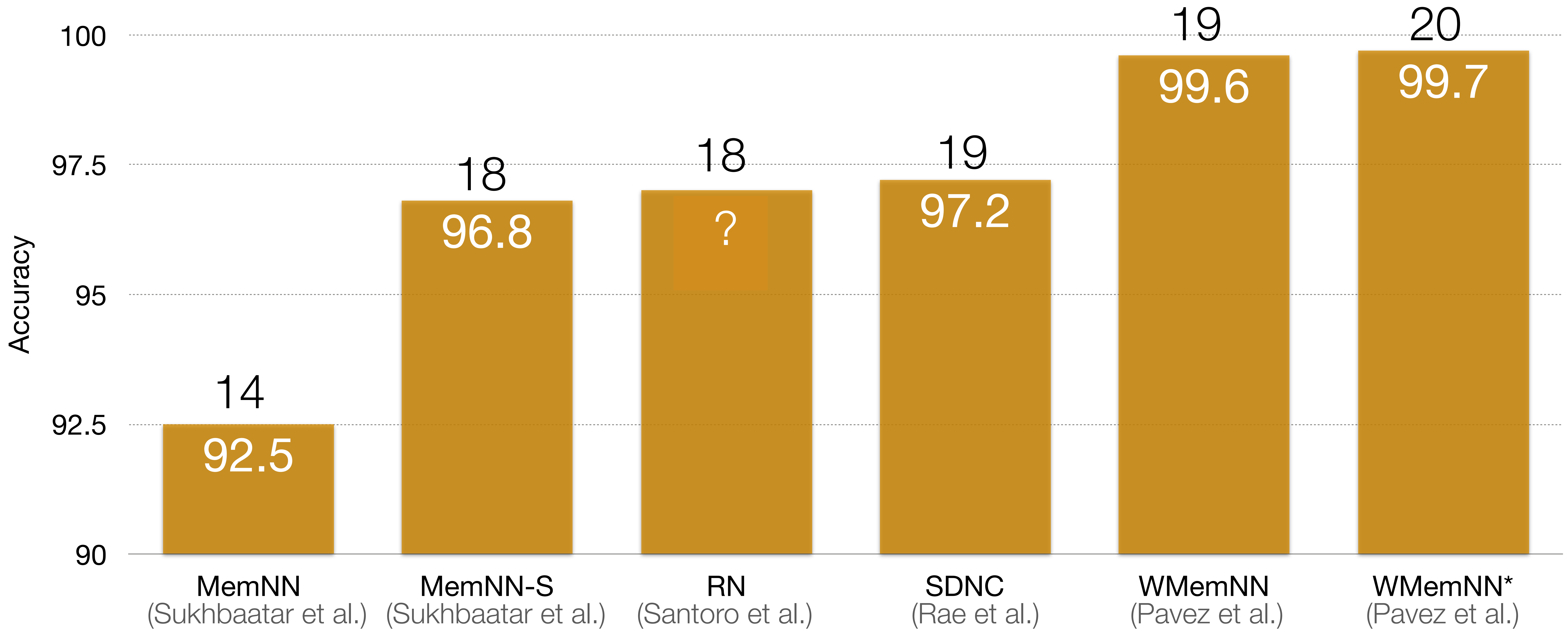
Results - Jointly trained bAbI-10k.

- Results on **jointly trained** bAbI-10k: Train a single model on all tasks simultaneously.
- Note that **EntNet** (Henaff et al.) solves all tasks in the **per-task version**: A single model for each task.



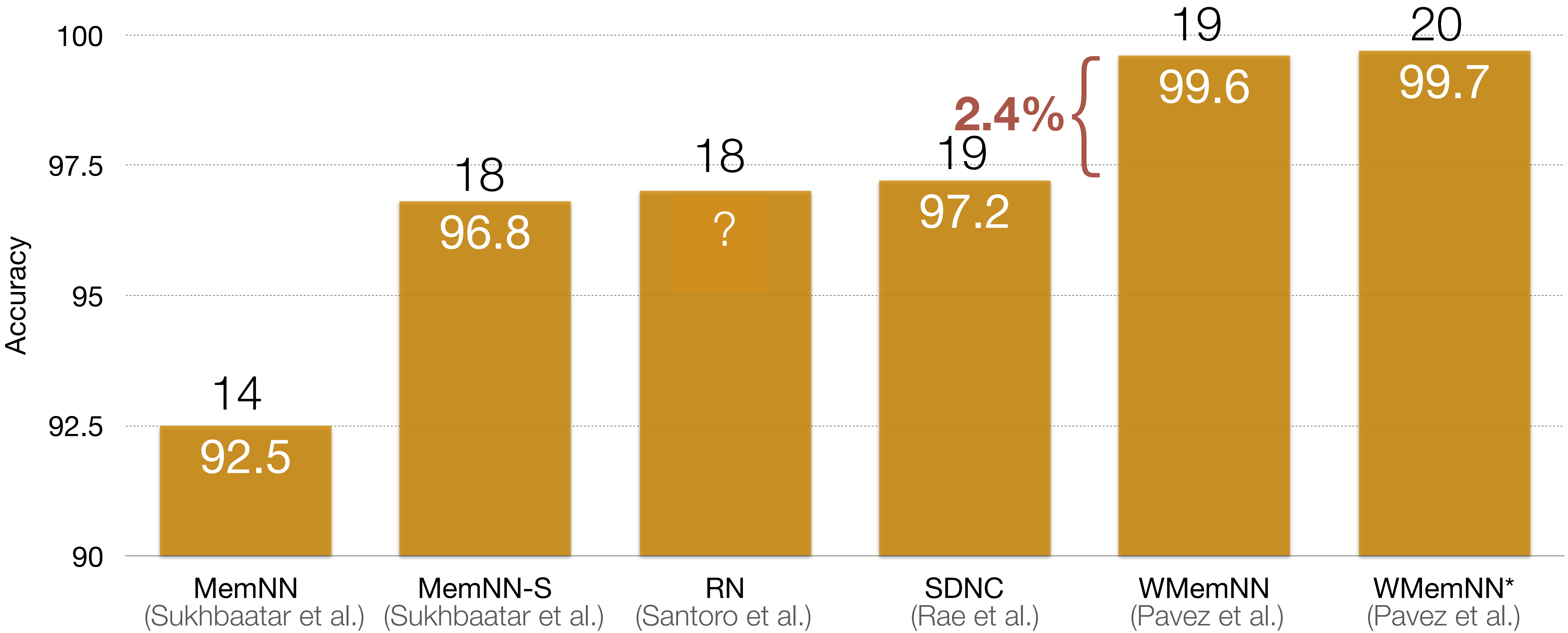
Results - Jointly trained bAbI-10k.

- Results on **jointly trained** bAbI-10k: Train a single model on all tasks simultaneously.
- Note that **EntNet** (Henaff et al.) solves all tasks in the **per-task version**: A single model for each task.



Results - Jointly trained bAbI-10k.

- Results on **jointly trained** bAbI-10k: Train a single model on all tasks simultaneously.
- Note that **EntNet** (Henaff et al.) solves all tasks in the **per-task version**: A single model for each task.



Ablations

complex attention patterns

multiple relations

	2 supporting facts	3 supporting facts	counting	basic induction	size reasoning	positional reasoning	path finding
MemNN (Sukhbaatar et al.)	99.0	93.2	94.4	99.2	92.0	59.2	25.3
MemNN(S) (Sukhbaatar et al.							
RN (Santoro et al.)							
WMemNN (no multi-head)							
WMemNN							

Ablations

complex attention patterns

multiple relations

	2 supporting facts	3 supporting facts	counting	basic induction	size reasoning	positional reasoning	path finding
MemNN (Sukhbaatar et al.)	99.0	93.2	94.4	99.2	92.0	59.2	25.3
MemNN(S) (Sukhbaatar et al.	100.0	99.7	96.7	100.0	97.9	73.4	69.1
RN (Santoro et al.)							
WMemNN (no multi-head)							
WMemNN							

Ablations

complex attention patterns

multiple relations

	2 supporting facts	3 supporting facts	counting	basic induction	size reasoning	positional reasoning	path finding
MemNN (Sukhbaatar et al.)	99.0	93.2	94.4	99.2	92.0	59.2	25.3
MemNN(S) (Sukhbaatar et al.	100.0	99.7	96.7	100.0	97.9	73.4	69.1
RN (Santoro et al.)	91.9	83.5		97.9			
WMemNN (no multi-head)							
WMemNN							

Ablations

complex attention patterns

multiple relations

	2 supporting facts	3 supporting facts	counting	basic induction	size reasoning	positional reasoning	path finding
MemNN (Sukhbaatar et al.)	99.0	93.2	94.4	99.2	92.0	59.2	25.3
MemNN(S) (Sukhbaatar et al.	100.0	99.7	96.7	100.0	97.9	73.4	69.1
RN (Santoro et al.)	91.9	83.5		97.9			
WMemNN (no multi-head)	98.6	90.3	98.8	50.2	99.9	99.7	97.2
WMemNN							

Ablations

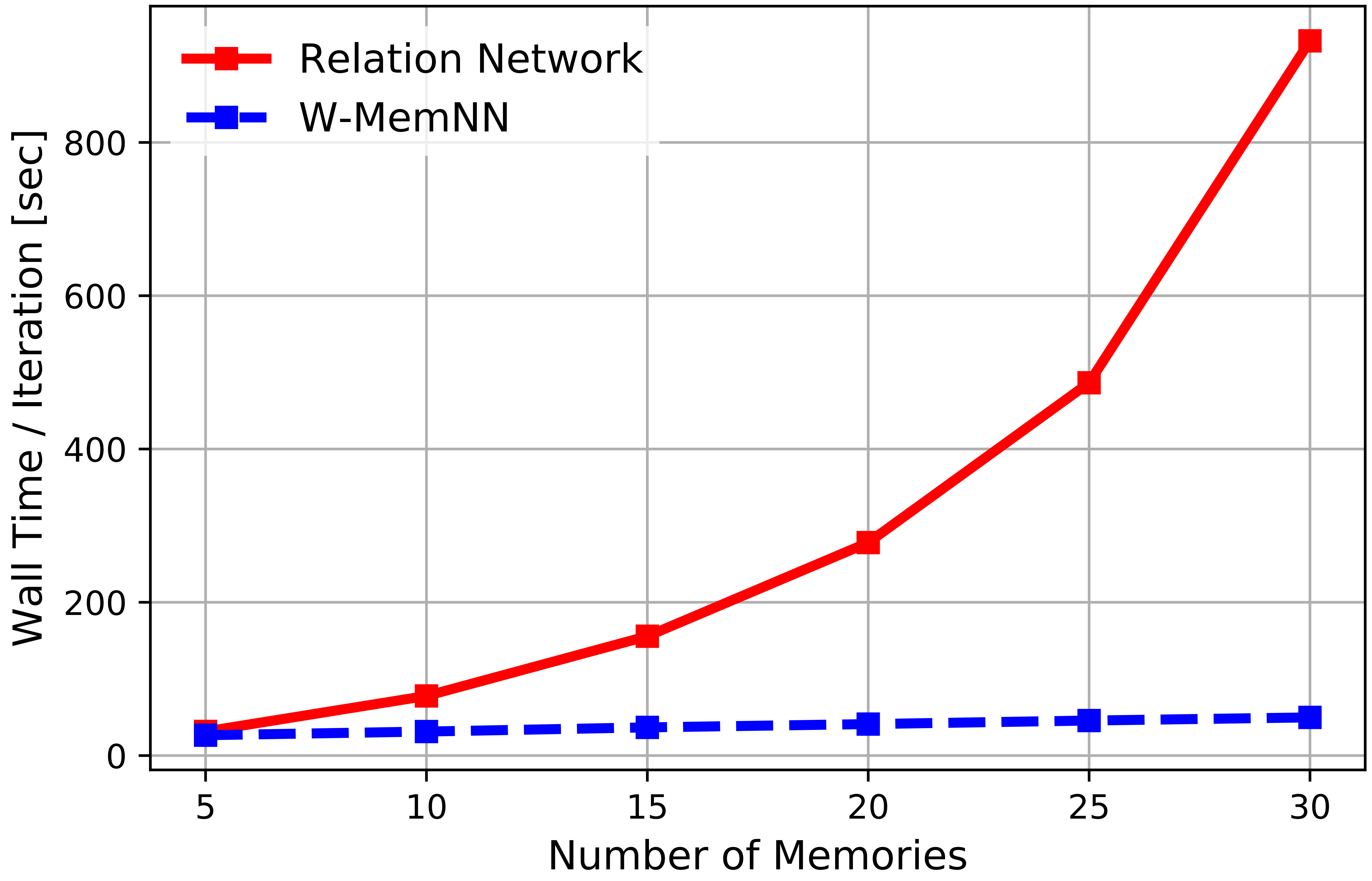
complex attention patterns

multiple relations

	2 supporting facts	3 supporting facts	counting	basic induction	size reasoning	positional reasoning	path finding
MemNN (Sukhbaatar et al.)	99.0	93.2	94.4	99.2	92.0	59.2	25.3
MemNN(S) (Sukhbaatar et al.	100.0	99.7	96.7	100.0	97.9	73.4	69.1
RN (Santoro et al.)	91.9	83.5		97.9			
WMemNN (no multi-head)	98.6	90.3	98.8	50.2	99.9	99.7	97.2
WMemNN	99.3	94.7	99.5	99.7	99.6	99.9	100.0

Time comparison

- Time comparisons for a forward and backward pass for a single batch of size 32.
- For 30 memories there is a speedup of almost 20x.



Conclusions

- We presented the **Working Memory Neural Network**, a Memory Network model augmented with a new **working memory buffer** and **relational reasoning module**.

- It retains the relational reasoning capabilities of the relation network while **reducing it computation times** considerably.

  - We hope that this contribution may help applying the relation network in **larger problems**.

Conclusions

- It is a very general **framework**. We argue that it should include:

Embedding + Short-term storage

Conclusions

- It is a very general **framework**. We argue that it should include:

Embedding + Short-term storage

Attentional controller + Working memory buffer

Conclusions

- It is a very general **framework**. We argue that it should include:

Embedding + Short-term storage

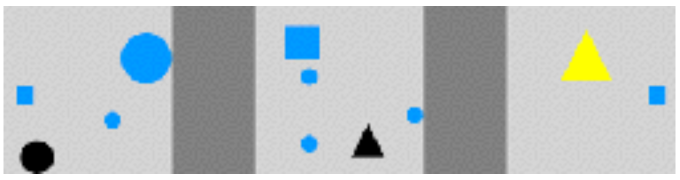
Attentional controller + Working memory buffer

Reasoning module

Conclusions

- It is a very general **framework**. We argue that it should include:

Embedding + Short-term storage



→ **CNN**

There is exactly one black triangle not touching any edge → **GRU**

Attentional controller + Working memory buffer

Multi-head attention

Reasoning module

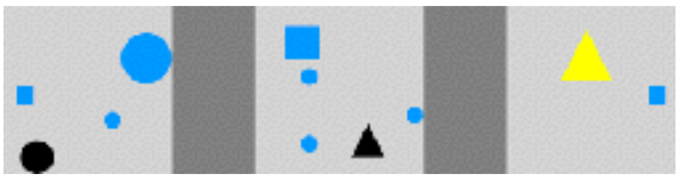
Relational Reasoning Module

NLVR: dev/test 65.6/65.8

Conclusions

- It is a very general **framework**. We argue that it should include:

Embedding + Short-term storage



CNN

There is exactly one black triangle not touching any edge

GRU

—-
16. Why, what are YOUR shoes done with?
17. said the Gryphon.
18. I mean, what makes then so shiny?
19. Alice looked down at then, and considered a little before she gave her answer.
20. They are done with blacking, I believe .

q. Boots and shoes under the sea. ' the ______ went in a deep voice, are done with a whiting.

biGRU

biGRU

Attentional controller + Working memory buffer

Multi-head attention

Scaled Dot-product Attention

Reasoning module

Relational Reasoning Module

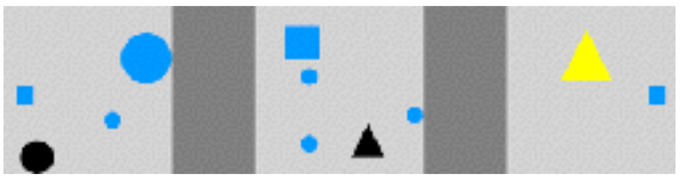
Attention Sum Reasoning Module

NLVR: dev/test 65.6/65.8

Conclusions

- It is a very general **framework**. We argue that it should include:

Embedding + Short-term storage



CNN

There is exactly one black triangle not touching any edge

GRU

—-
16. Why, what are YOUR shoes done with?
17. said the Gryphon.
18. I mean, what makes then so shiny?
19. Alice looked down at then, and considered a little before she gave her answer.
20. They are done with blacking, I believe .

q. Boots and shoes under the sea. ‘ the ______ went in a deep voice, are done with a whiting.

biGRU

biGRU

Elmo LSTM ResNet

Attentional controller + Working memory buffer

Multi-head attention

Scaled Dot-product Attention

BiAtt AoA GA

Reasoning module

Relational Reasoning Module

Attention Sum Reasoning Module

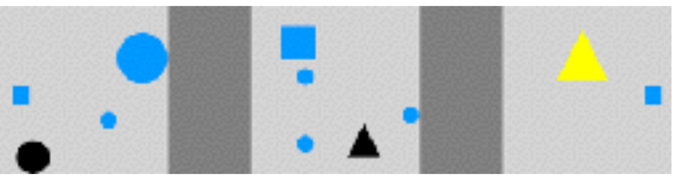
Attention Sum Decoder Graph NN

NLVR: dev/test 65.6/65.8

Conclusions

- It is a very general **framework**. We argue that it should include:

Embedding + Short-term storage



CNN

There is exactly one black triangle not touching any edge

GRU

—-
16. Why, what are YOUR shoes done with?
17. said the Gryphon.

```
def input_module(x, u, adjacent=None, use_lstm=False, seq_lstm=
    '''
      Process input and create memories to be stored in short-t
    '''
    if adjacent == None:
        layer_encoder_m = Embedding(input_dim=vocab_size,
                                    output_dim=EMBED_HIDDEN_SIZE,
                                    input_length=story_maxlen,
                                    embeddings_initializer='gloro

        if use_lstm:
```

Attentional controller + Working memory buffer

Multi-head attention

```
def attention_module(memories, u, use_softmax=True, MLP=None
    '''
    Multi-head attention mechanism implementation
    '''

    head_outs = []
    mems = []
    for k in range(n_heads):
        # results are similar for linear or tanh activation
        head = TimeDistributed(Dense(LSTM_HIDDEN_UNITS,  ker
                               activation='tanh', use_
```

Reasoning module

Relational Reasoning Module

NLVR:
dev/test
65.6/65.8

```
def reasoning_module(working_buffer):
    '''
    Implementation of the relational reasoning module
    '''

    relations = []
    for fact_object_1 in working_buffer:
        for fact_object_2 in working_buffer:
            relations.append(concatenate([fact_object_1, fact_objec
```

Thanks!

juan.pavezs@alumnos.usm.cl

Code: https://github.com/jgpavez/Working-Memory-Networks

@juanpavez