# Multimodal Deep Learning

Russ Salakhutdinov

Machine Learning Department
Carnegie Mellon University
Canadian Institute for Advanced Research

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

# Talk Roadmap

- Caption Generation with Multiplicative Neural Language Model
- Image Generation from Captions with Attention
- Open Problems and Research Directions

# Example: Understanding Images

TAGS:

strangers, coworkers, conventioneers, attendants, patrons

Nearest Neighbor Sentence:

people taking pictures of a crazy person

Model Samples

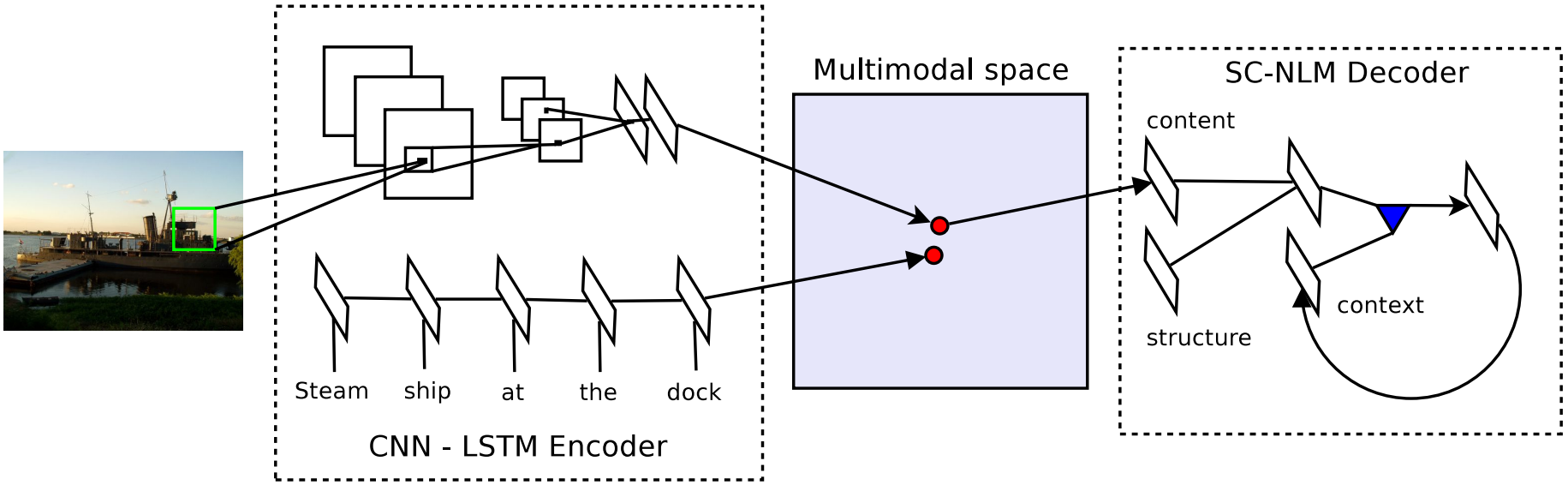
- a group of people in a crowded area
- a group of people are walking and talking .
- a group of people, standing around and talking .

# Generating Sentences

- More challenging problem.
- How can we generate complete descriptions of images?

| Input | Output |
| --- | --- |
|  | A man skiing down the snow covered mountain with a dark sky in the background. |

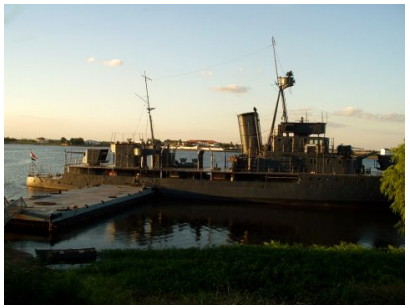
# Encode-Decode Framework

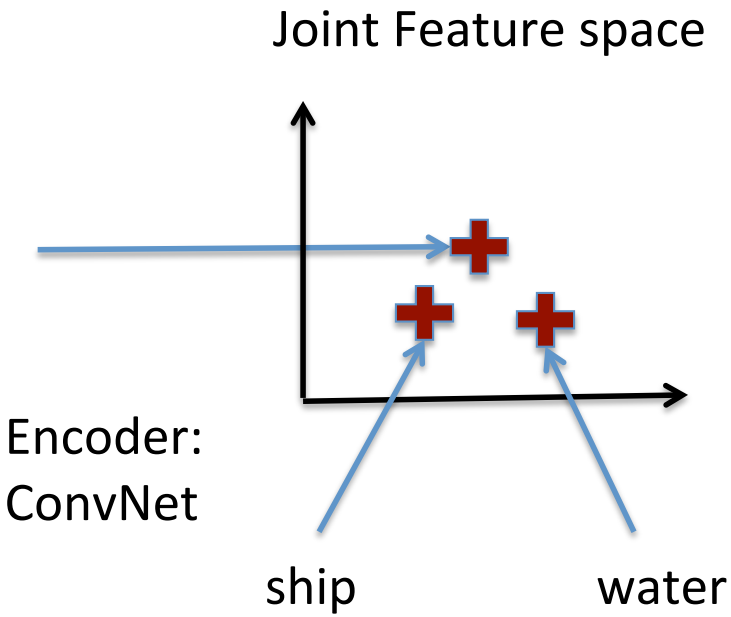



• Encoder: CNN and Recurrent Neural Net for a joint image-sentence embedding.

• Decoder: A neural language model for generating a sequence of words.

# An Image-Text Encoder





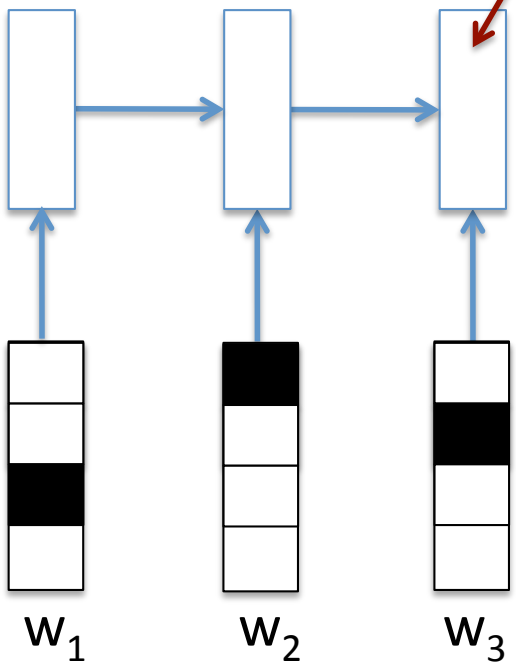

- Learn a joint embedding space of images and text:
  - Can condition on anything (images, words, phrases, etc)
  - Natural definition of a scoring function (inner products in the joint space).

Socher 2013, Frome 2013, Kiros 2014

# An Image-Text Encoder

Sentence Representation

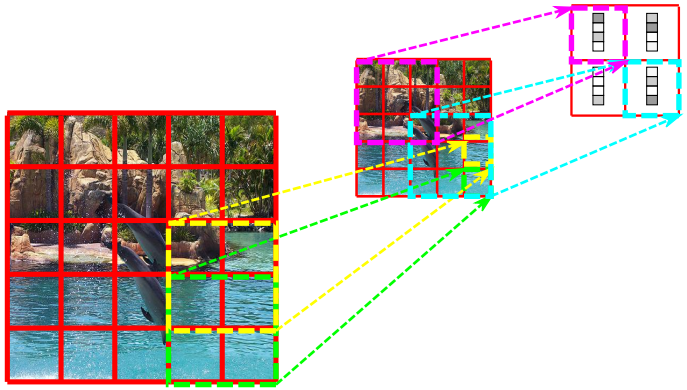
See Skip-Thought Vectors, (Kiros et.al. arXiv 2015)




1-of-V encoding of words

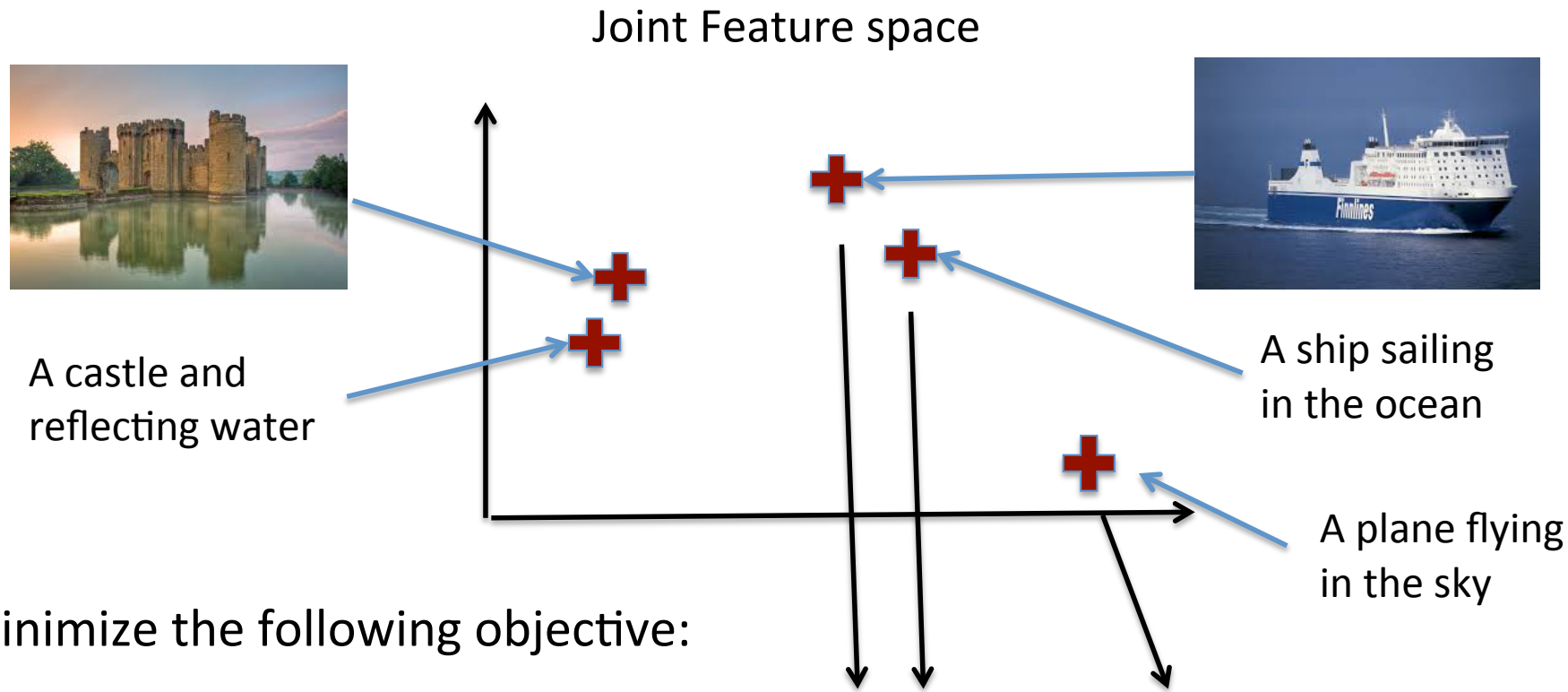
Recurrent Neural Network (LSTM)

Image Representation



Convolutional Neural Network

# An Image-Text Encoder




Minimize the following objective:

Images: $$\sum_{\mathbf{x}} \sum_{k} \max\{0, \alpha - s(\mathbf{x}, \mathbf{v}) + s(\mathbf{x}, \mathbf{v}_k)\} +$$

Text: $$\sum_{\mathbf{v}} \sum_{k} \max\{0, \alpha - s(\mathbf{v}, \mathbf{x}) + s(\mathbf{v}, \mathbf{x}_k)\}$$

(Kiros, Salakhutdinov, Zemel, TACL 2015)

# Retrieving Sentences for Images

The dogs are in the snow in front of a fence .


Four men playing basketball , two from each team .

A boy skateboarding

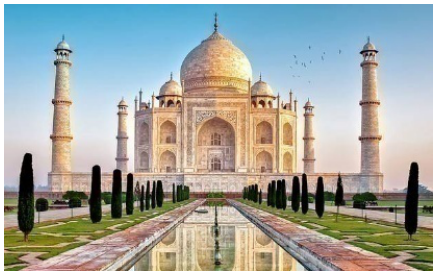
Two men and a woman smile at the camera .


Women participate in a skit onstage .

A man is doing tricks on a bicycle on ramps in front of a crowd .

(Kiros, Salakhutdinov, Zemel, TACL 2015)

# Tagging and Retrieval



mosque, tower, building, cathedral, dome, castle

ski, skiing, skiers, skiiers, snowmobile

kitchen, stove, oven, refrigerator, microwave

bowl, cup, soup, cups, coffee

beach

snow

# Retrieval with Adjectives

fluffy

delicious

(Kiros, Salakhutdinov, Zemel, TACL 2015)

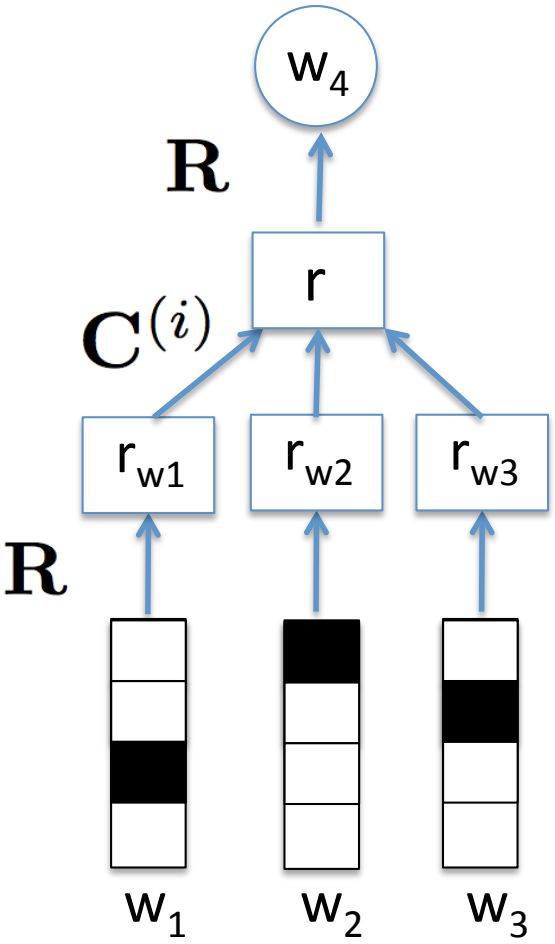
# How About Generating Sentences!

Input

Output

A man skiing down the snow covered mountain with a dark sky in the background.

# Log-bilinear Neural Language Model



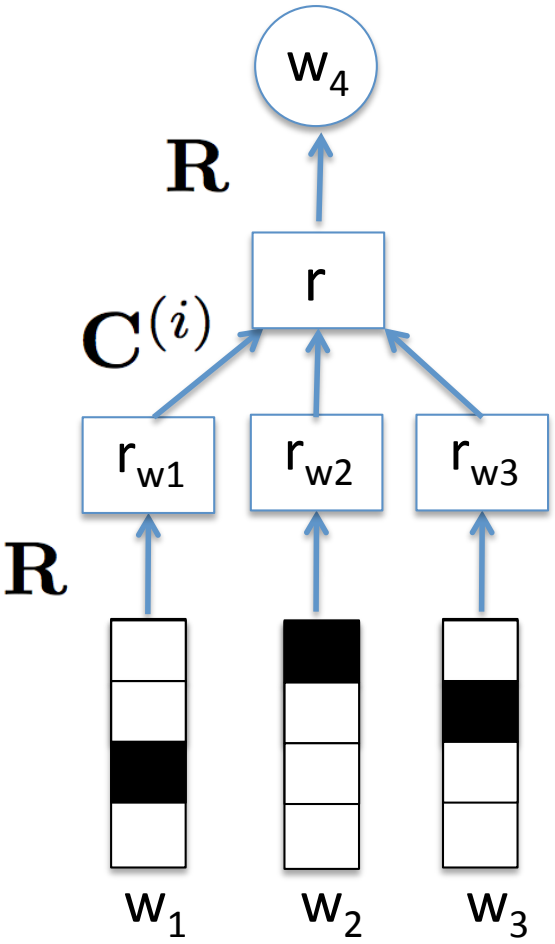

1-of-V encoding of words

- Feedforward neural network with a single linear hidden layer.
- Each word w is represented as a K-dim real-valued vector $r_w \in R^K$.
- **R** denote the V × K matrix of word representation vectors, where V is the vocabulary size.
- $(w_1, \ldots, w_{n-1})$ is tuple of n-1 words, where n-1 is the context size. The next word representation becomes:

$$\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \mathbf{C}^{(i)} \mathbf{r}_{w_i},$$

K × K context parameter matrices

# Log-bilinear Neural Language Model




1-of-V encoding of words

$$\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \mathbf{C}^{(i)} \mathbf{r}_{w_i},$$

Predicted representation of $r_{wn}$.

- The conditional probability of the next word given by:

$$P(w_n = i | w_{1:n-1}) = \frac{\exp(\hat{\mathbf{r}}^T \mathbf{r}_i + b_i)}{\sum_{j=1}^{V} \exp(\hat{\mathbf{r}}^T \mathbf{r}_j + b_j)}$$

Can be expensive to compute

Bengio et.al. 2003

# Multiplicative Model

- We represent words as a tensor:

$$\boldsymbol{\mathcal{T}} \in \mathbb{R}^{V \times K \times G}$$

where G is the number of tensor slices.

- Given an attribute vector $u \in R^G$ (e.g. image features), we can compute attribute-gated word representations as:

$$\boldsymbol{\mathcal{T}}^u = \sum_{i=1}^{G} u_i \boldsymbol{\mathcal{T}}^{(i)}$$

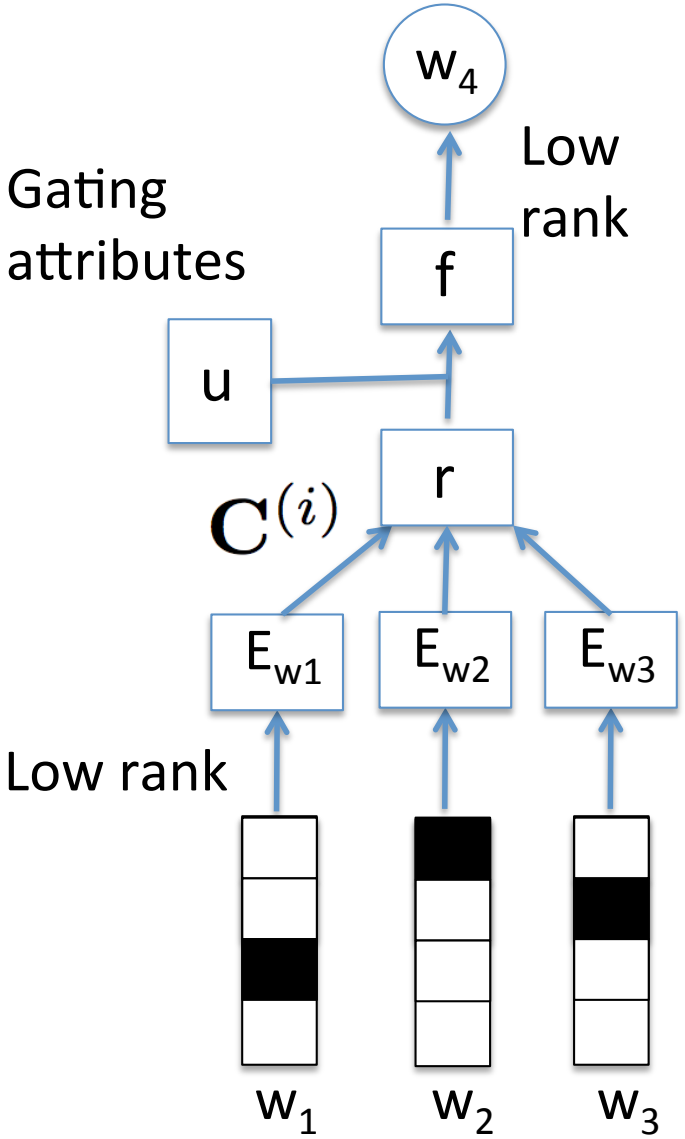
- Re-represent Tensor in terms of 3 lower-rank matrices (where F is the number of pre-chosen factors):

$$\mathbf{W}^{fk} \in \mathbb{R}^{F \times K}, \mathbf{W}^{fd} \in \mathbb{R}^{F \times G} \quad \mathbf{W}^{fv} \in \mathbb{R}^{F \times V}$$

$$\boldsymbol{\mathcal{T}}^u = (\mathbf{W}^{fv})^\top \cdot \mathrm{diag}(\mathbf{W}^{fd}\mathbf{u}) \cdot \mathbf{W}^{fk}$$

(Kiros, Zemel, Salakhutdinov, NIPS 2014)

# Multiplicative Log-bilinear Model




- Let $\mathbf{E} = (\mathbf{W}^{fk})^\top \mathbf{W}^{fv}$ denote a folded K $\times$ V matrix of word embeddings.
- Then the predicted next word representation is:

$$\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \mathbf{C}^{(i)} \mathbf{E}(:, w_i)$$

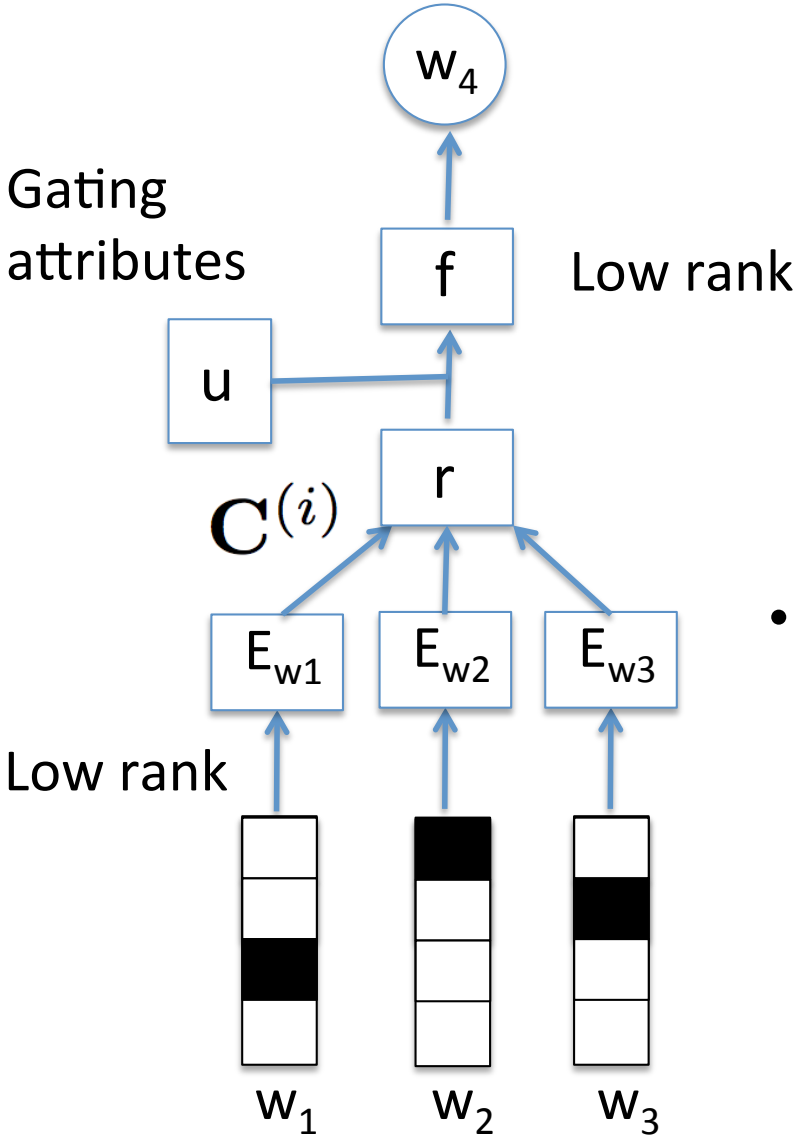
- Given next word representation r, the factor outputs are:

$$\mathbf{f} = (\mathbf{W}^{fk} \hat{\mathbf{r}}) \bullet (\mathbf{W}^{fd} \mathbf{x})$$

Component-wise product

(Kiros, Zemel, Salakhutdinov, NIPS 2014)

# Multiplicative Log-bilinear Model




$$\mathbf{E} = (\mathbf{W}^{fk})^\top \mathbf{W}^{fv}$$

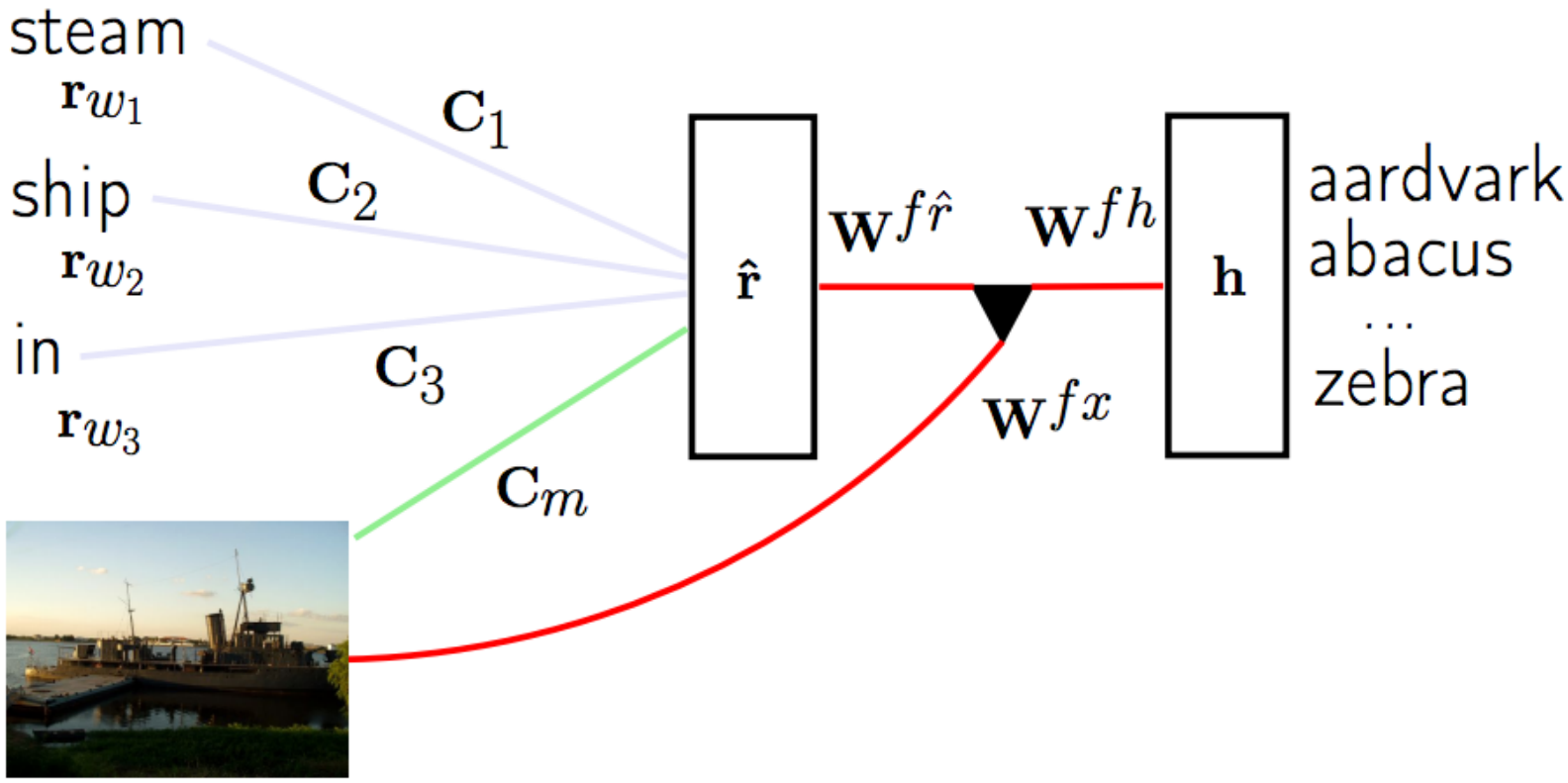
$$\hat{\mathbf{r}} = \sum_{i=1}^{n-1} \mathbf{C}^{(i)} \mathbf{E}(:, w_i)$$

$$\mathbf{f} = (\mathbf{W}^{fk} \hat{\mathbf{r}}) \bullet (\mathbf{W}^{fd} \mathbf{x})$$

- The conditional probability of the next word is given by:

$$P(w_n = i | w_{1:n-1}, \mathbf{u}) = \frac{\exp\left((\mathbf{W}^{fv}(:, i))^\top \mathbf{f} + b_i\right)}{\sum_{j=1}^{V} \exp\left((\mathbf{W}^{fv}(:, j))^\top \mathbf{f} + b_j\right)}$$

# Decoding: Neural Language Model




- Image features are gating the hidden-to-output connections when predicting the next word!
- We can also condition on POS tags when generating a sentence.

(Kiros, Salakhutdinov, Zemel, ICML 2014)

# Decoding: Structured NLM

_____ (NN VBN IN DT NN)

DT

$$P(w_n | w_{1:n-1}, t_{n:n+k}, \mathbf{x})$$

n-th word  word context  POS context

# Decoding: Structured NLM

_____ (NN VBN IN DT NN)
DT

A _____ (VBN IN DT NN -)
NN

$$P(w_n | w_{1:n-1}, t_{n:n+k}, \mathbf{x})$$

n-th word  word context  POS context

# Decoding: Structured NLM

$$P(w_n | w_{1:n-1}, t_{n:n+k}, \mathbf{x})$$

n-th word  word context  POS context

_____ (NN VBN IN DT NN)
DT

A _____ (VBN IN DT NN -)
NN

A bicycle _____ (IN DT NN - -)
VBN

# Decoding: Structured NLM

$$P(w_n | w_{1:n-1}, t_{n:n+k}, \mathbf{x})$$

n-th word  word context  POS context

_____ (NN VBN IN DT NN)
DT

A _____ (VBN IN DT NN -)
NN

A bicycle _____ (IN DT NN - -)
VBN

A bicycle parked _____ (DT NN - - -)
IN

# Decoding: Structured NLM

$$P(w_n | w_{1:n-1}, t_{n:n+k}, \mathbf{x})$$

n-th word  word context  POS context

_____ (NN VBN IN DT NN)
DT

A _____ (VBN IN DT NN -)
NN

A bicycle _____ (IN DT NN - -)
VBN

A bicycle parked _____ (DT NN - - -)
IN

A bicycle parked on _____ (NN - - - -)
DT

A bicycle parked on the _____ (- - - - -)
NN

# Caption Generation

a car is parked in
the middle of nowhere .

a wooden table and chairs
arranged in a room .

there is a cat sitting on a shelf .

a ferry boat on a marina
with a group of people .

a little boy with a bunch
of friends on the street .

(Kiros, Salakhutdinov, Zemel, TACL 2015)

# Caption Generation

the two birds are trying to be seen in the water . (can't count)

a giraffe is standing next to a fence in a field . (hallucination)

a parked car while driving down the road . (contradiction)

# Caption Generation

the two birds are trying to be seen in the water . (can't count)

a giraffe is standing next to a fence in a field . (hallucination)

a parked car while driving down the road . (contradiction)

the handlebars are trying to ride a bike rack . (nonsensical)

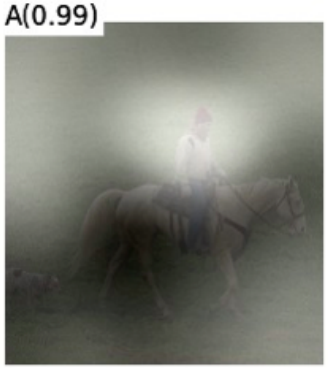
a woman and a bottle of wine in a garden . (gender)

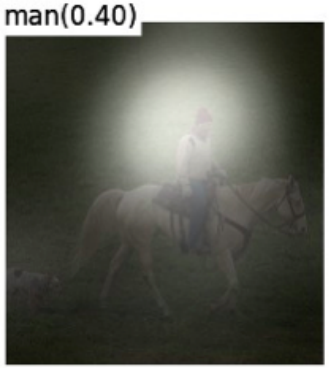
# Caption Generation

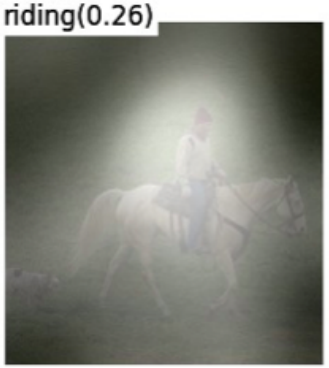
A man holding a red apple in his mouth

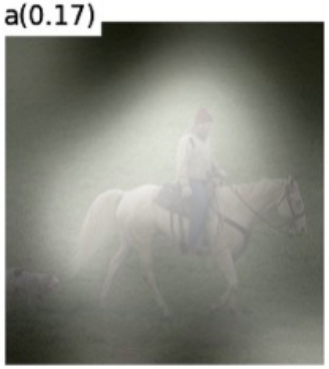

Two men are playing frisbee in a park area

# Caption Generation with Visual Attention

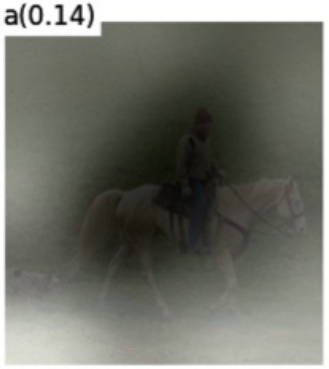







A man riding a horse in a field.

Xu et.al., ICML 2015

# Neural Story Telling

**Sample from the Generative Model (recurrent neural network):**

She was in love with him for the first time in months, so she had no intention of escaping.

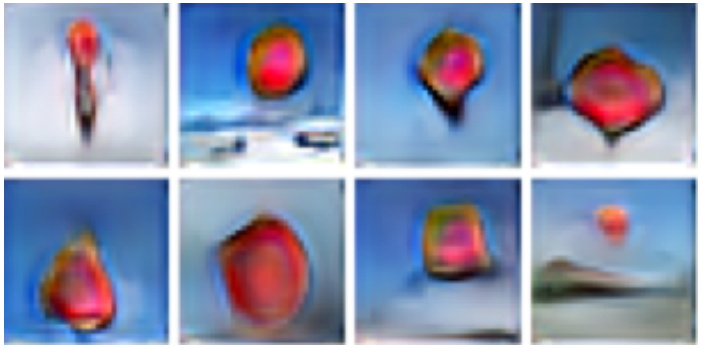
The sun had risen from the ocean, making her feel more alive than normal. She is beautiful, but the truth is that I do not know what to do. The sun was just starting to fade away, leaving people scattered around the Atlantic Ocean.

(Kiros et al., NIPS 2015)

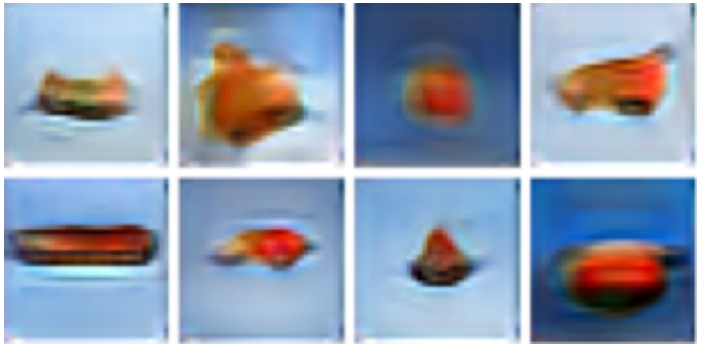
# Generating Images from Captions

- Can we generate images from natural language descriptions?

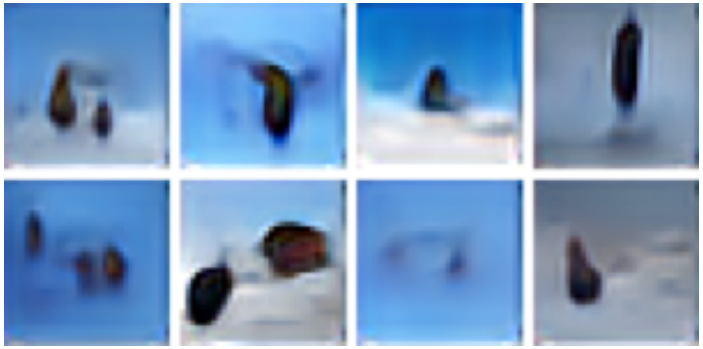
A **stop sign** is flying in blue skies



A **pale yellow school bus** is flying in blue skies



A **herd of elephants** is flying in blue skies



A **large commercial airplane** is flying in blue skies

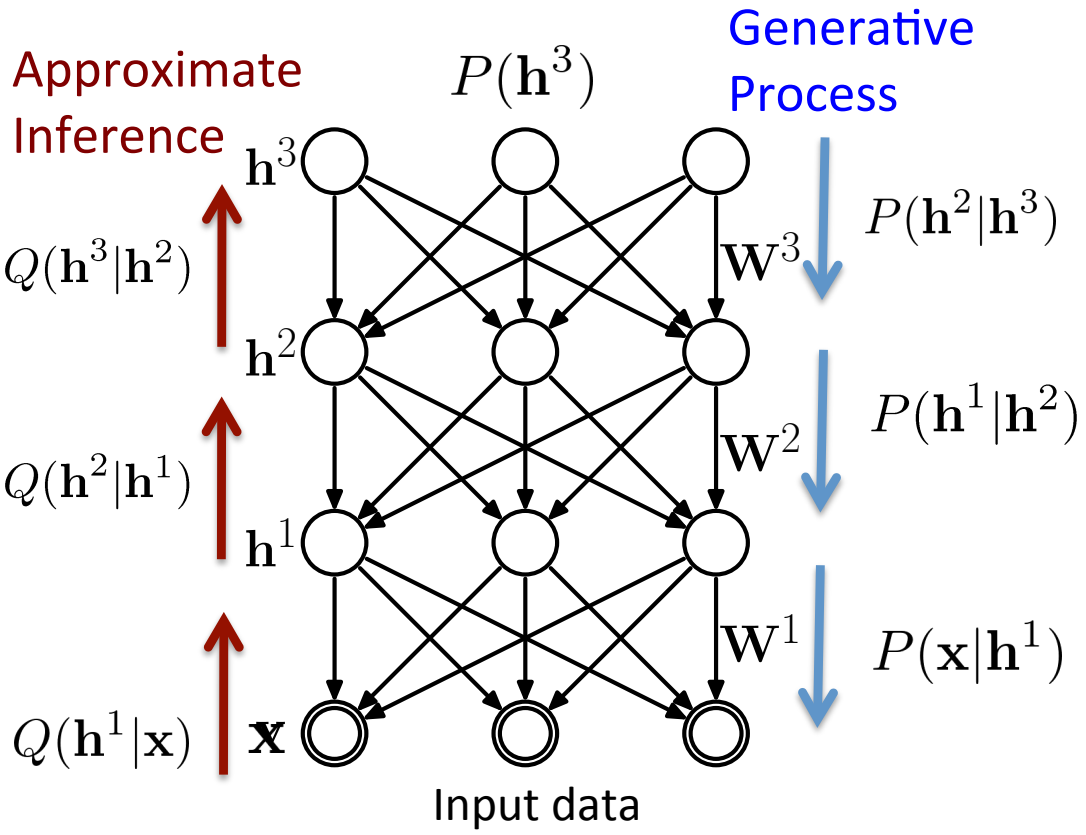
(Mansimov, Parisotto, Ba, Salakhutdinov, ICLR 2016)

# Talk Roadmap

- Caption Generation with Multiplicative Neural Language Model
- Image Generation from Captions with Attention
- Open Problems and Research Directions

# Helmholtz Machines

- Hinton, G. E., Dayan, P., Frey, B. J. and Neal, R., Science 1995



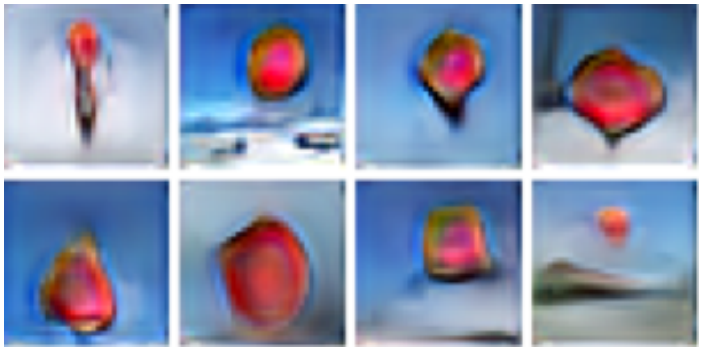

- Kingma & Welling, 2014
- Rezende, Mohamed, Daan, 2014
- Mnih & Gregor, 2014
- Bornschein & Bengio, 2015
- Gregor et. al., 2015
- Tang & Salakhutdinov, 2013

# Motivating Example

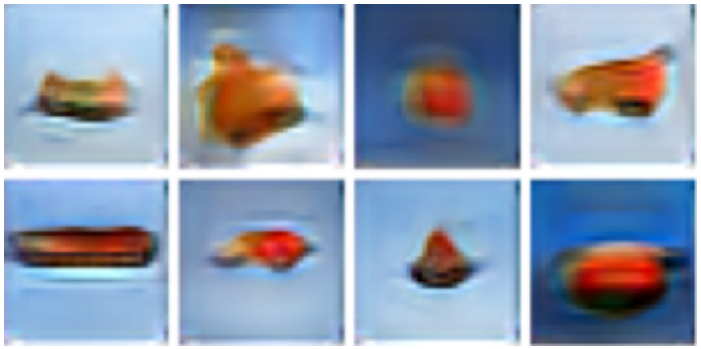
(Mansimov, Parisotto, Ba, Salakhutdinov, 2015)

- Can we generate images from natural language descriptions?

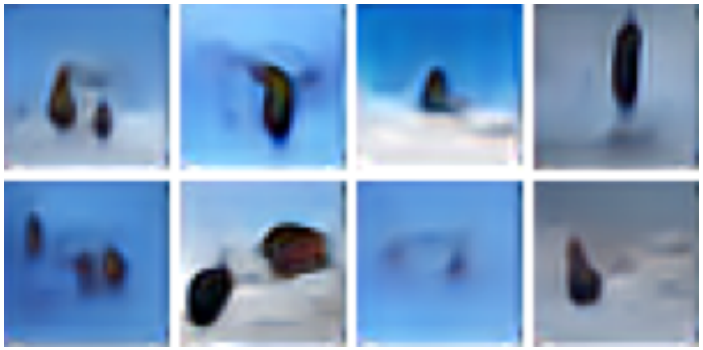
A **stop sign** is flying in blue skies



A **pale yellow school bus** is flying in blue skies



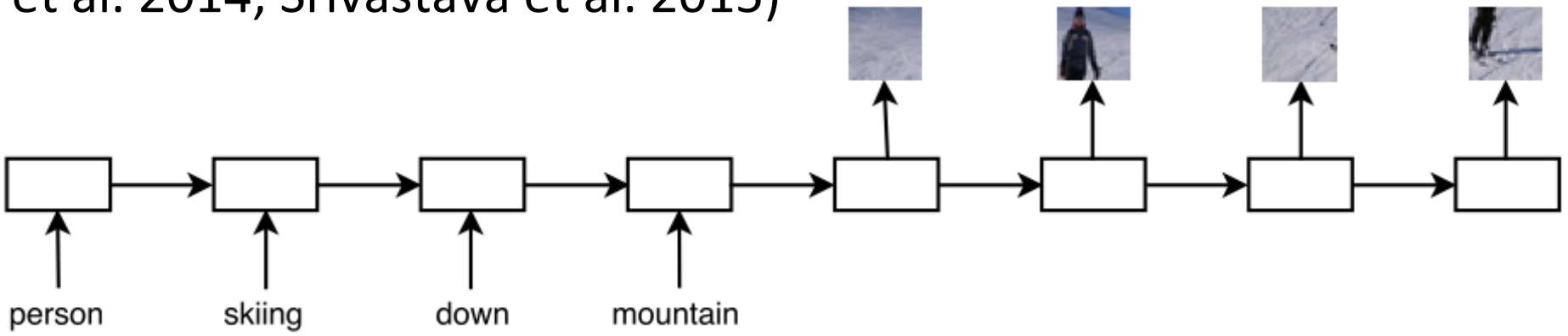
A **herd of elephants** is flying in blue skies



A **large commercial airplane** is flying in blue skies

# Sequence-to-Sequence

• Sequence-to-sequence framework. (Sutskever et al. 2014; Cho et al. 2014; Srivastava et al. 2015)



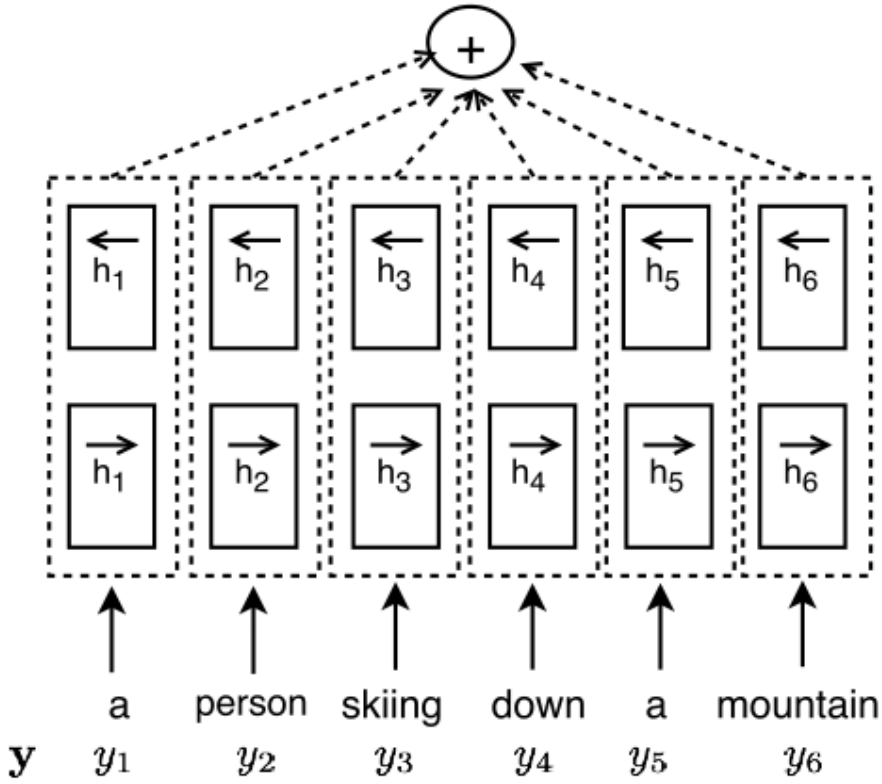

• Caption (**y**) is represented as a sequence of consecutive words.

• Image (**x**) is represented as a sequence of patches drawn on canvas.

• Attention mechanism over:

- Words: Which words to focus on when generating a patch
- Image Location Where to place the generated patches on the canvas

# Representing Captions
# Bidirectional RNN




- Forward RNN reads the sentence **y** from left to right:

$$[\overrightarrow{\mathbf{h}}_1^{lang}, \overrightarrow{\mathbf{h}}_2^{lang}, , \overrightarrow{\mathbf{h}}_N^{lang}]$$

- Backward RNN reads the sentence **y** from right to left:

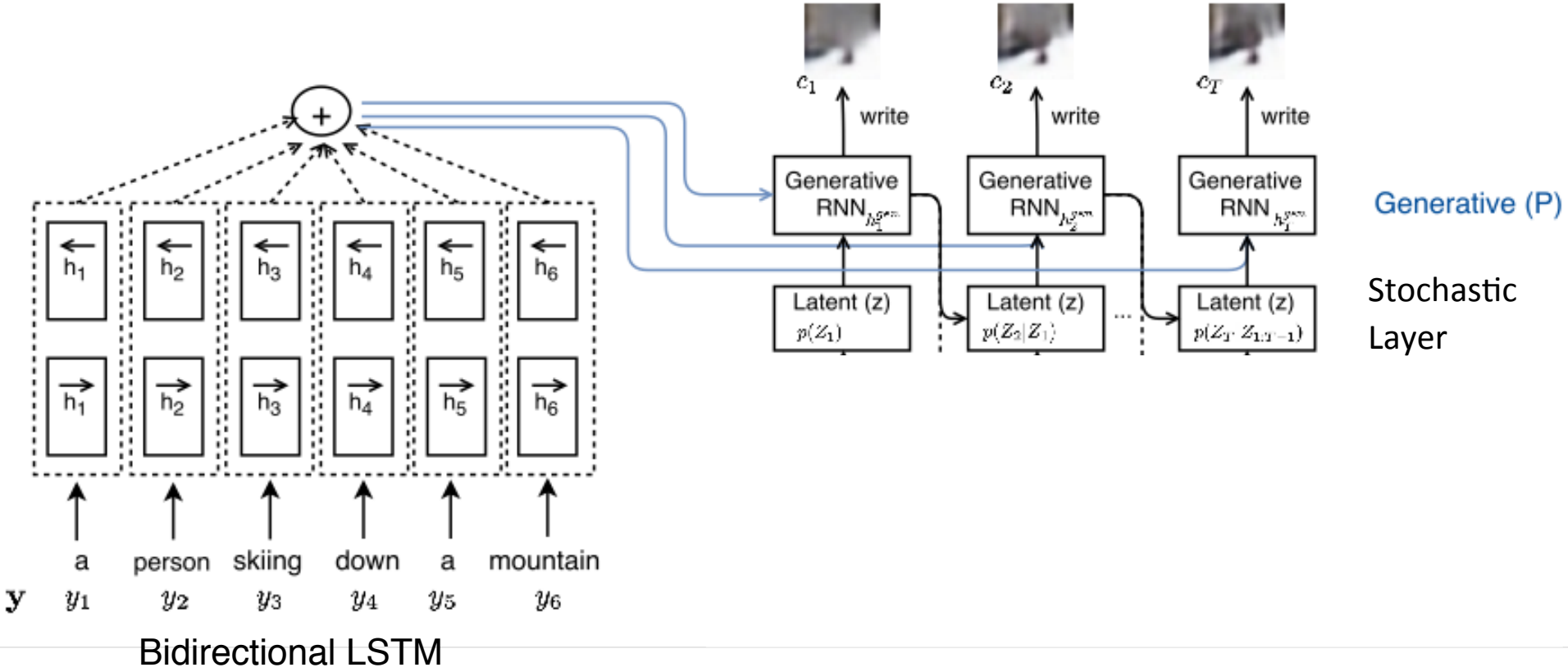
$$[\overleftarrow{\mathbf{h}}_1^{lang}, \overleftarrow{\mathbf{h}}_2^{lang}, , \overleftarrow{\mathbf{h}}_N^{lang}]$$

- The hidden states are then concatenated:

$$\mathbf{h}^{lang} = [\mathbf{h}_1^{lang}, \mathbf{h}_2^{lang}, \ldots, \mathbf{h}_N^{lang}], \quad \text{with } \mathbf{h}_i^{lang} [\overrightarrow{\mathbf{h}}_i^{lang}, \overleftarrow{\mathbf{h}}_i^{lang}]$$

# Overall Model

(Mansimov, Parisotto, Ba, Salakhutdinov, 2015)

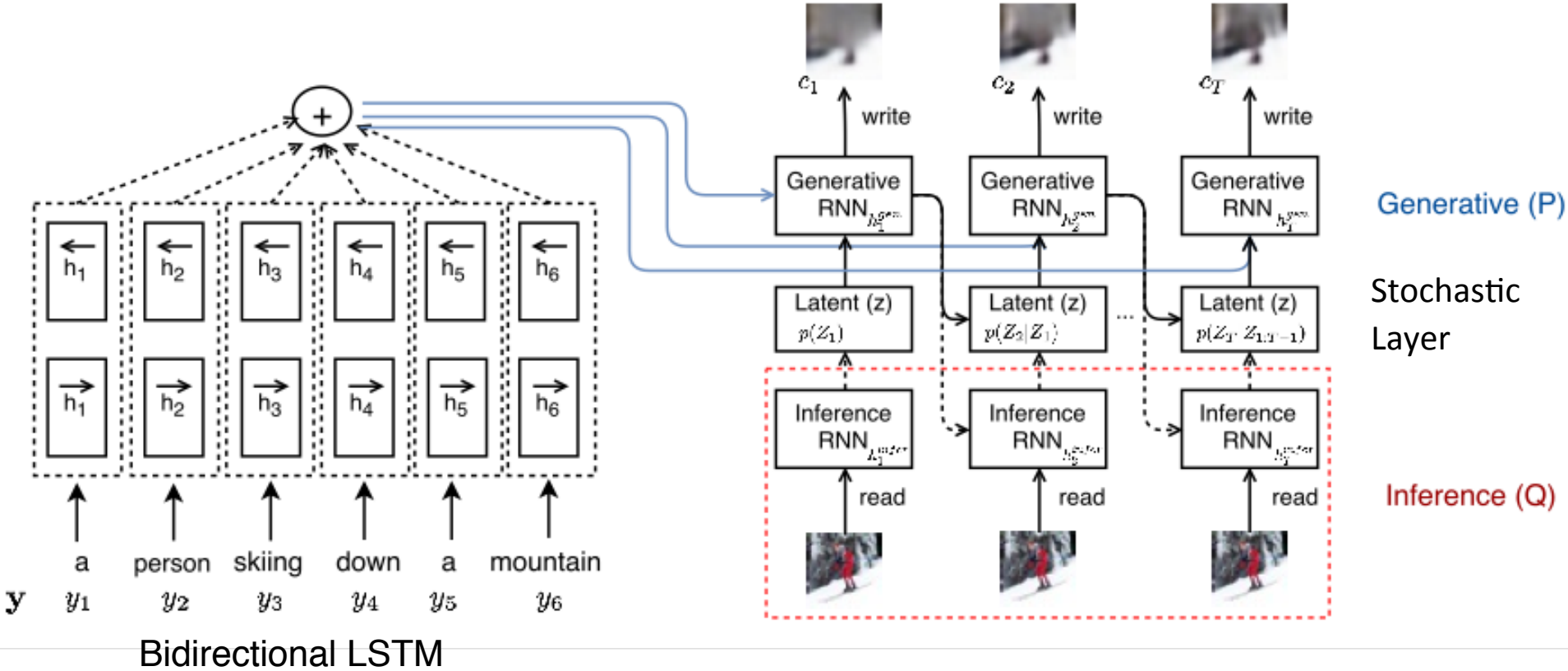



- Generative Model: Stochastic Recurrent Network, chained sequence of Variational Autoencoders, with a single stochastic layer.

Gregor et. al. 2015

# Overall Model

(Mansimov, Parisotto, Ba, Salakhutdinov, 2015)

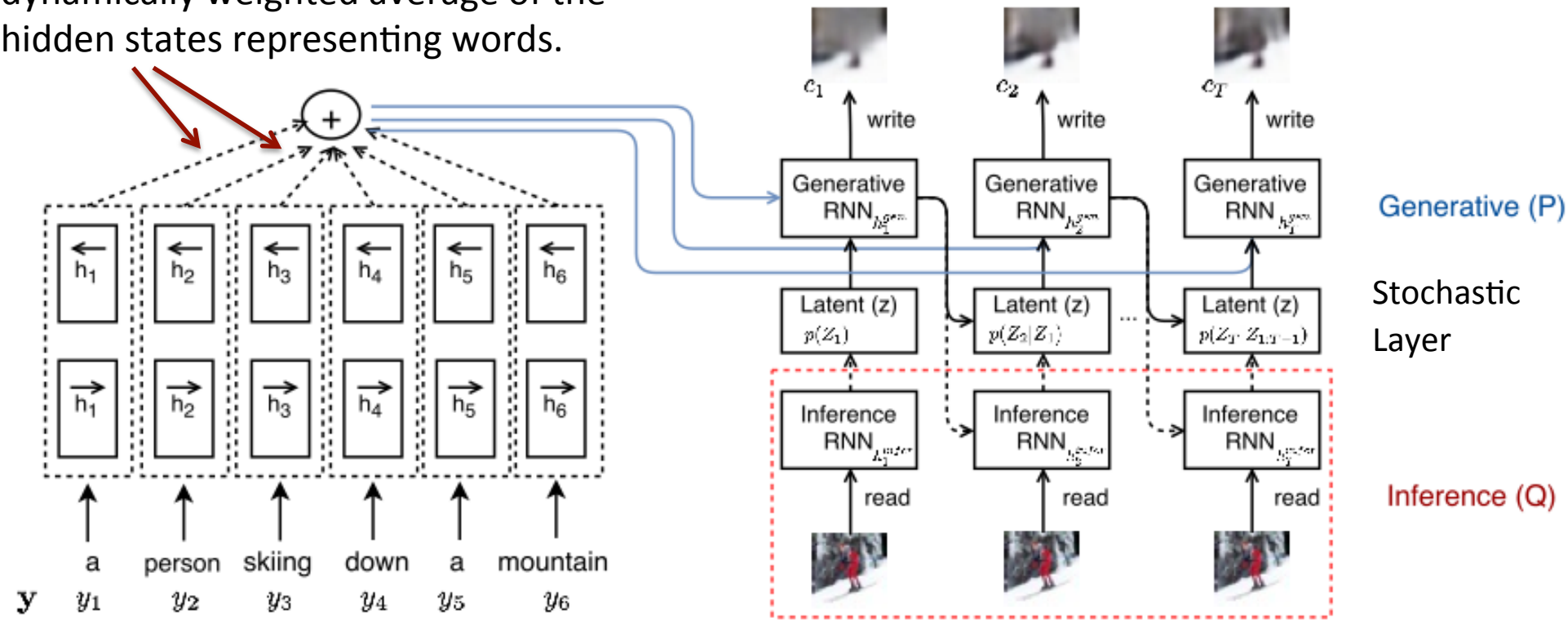



- Generative Model: Stochastic Recurrent Network, chained sequence of Variational Autoencoders, with a single stochastic layer.
- Recognition Model: Deterministic Recurrent Network.

Gregor et. al. 2015

# Overall Model

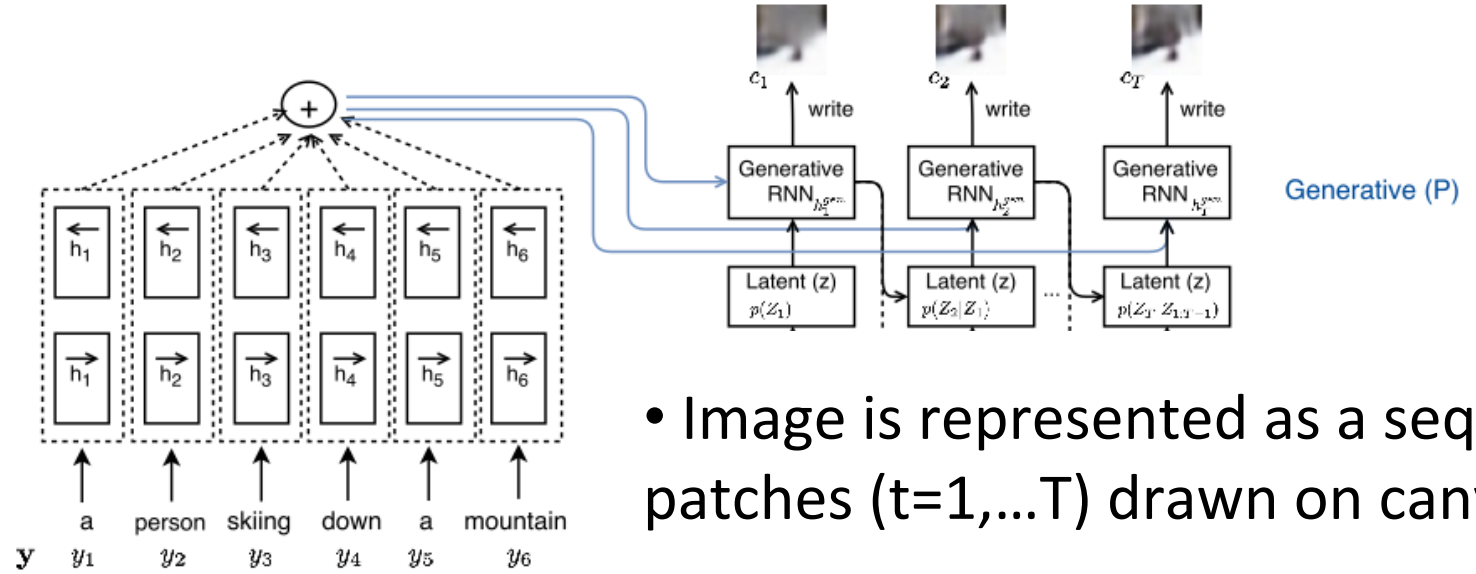
Sentence representation:
dynamically weighted average of the hidden states representing words.




- Attention (alignment): Focus on different words at different time steps when generating patches and placing them on the canvas.

Bahdanau et. al. 2015

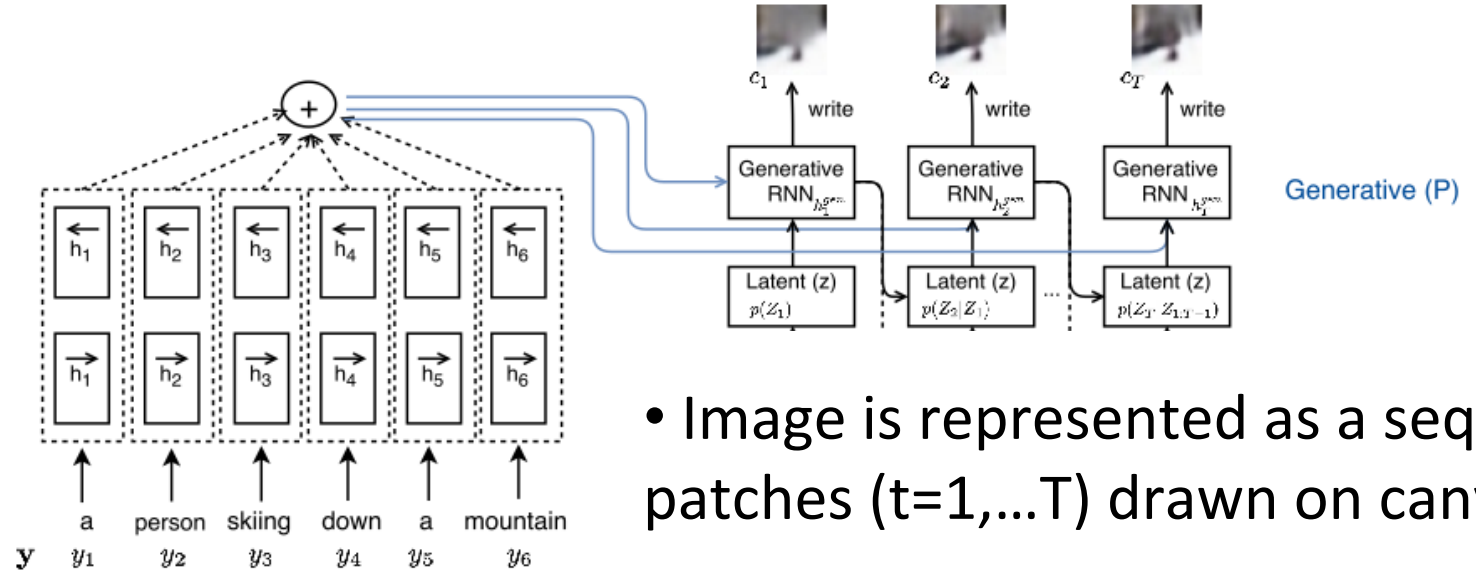
# Generating Images




- Image is represented as a sequence of patches (t=1,…T) drawn on canvas:

$$\mathbf{z}_t \sim P(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}) = \mathcal{N}\big(\mu(\mathbf{h}_{t-1}^{gen}), \sigma(\mathbf{h}_{t-1}^{gen})\big), \quad P(\mathbf{Z}_1) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

# Generating Images

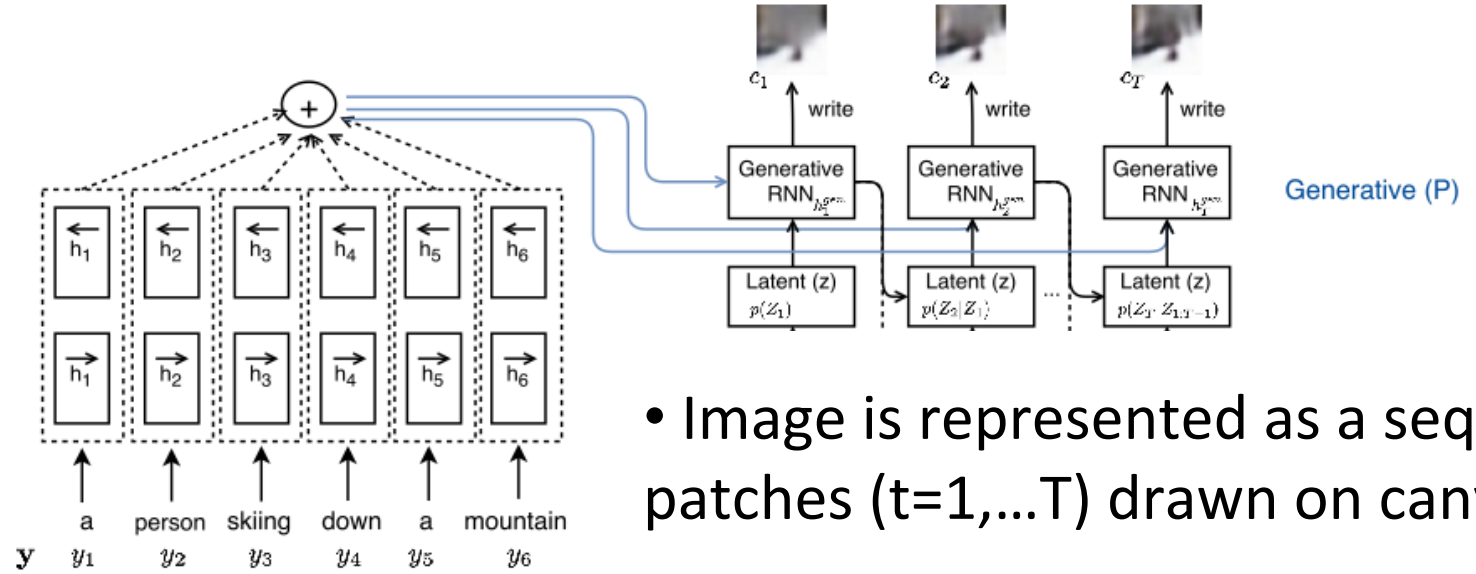



- Image is represented as a sequence of patches (t=1,…T) drawn on canvas:

$$\mathbf{z}_t \sim P(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}) = \mathcal{N}\big(\mu(\mathbf{h}_{t-1}^{gen}), \sigma(\mathbf{h}_{t-1}^{gen})\big), \quad P(\mathbf{Z}_1) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

$$s_t = align\big(\mathbf{h}_{t-1}^{gen}, \mathbf{h}^{lang}\big) \quad \mathbf{h}_t^{gen} = LSTM^{gen}\big(\mathbf{h}_{t-1}^{gen}, [\mathbf{z}_t, s_t]\big)$$

# Generating Images




- Image is represented as a sequence of patches (t=1,…T) drawn on canvas:

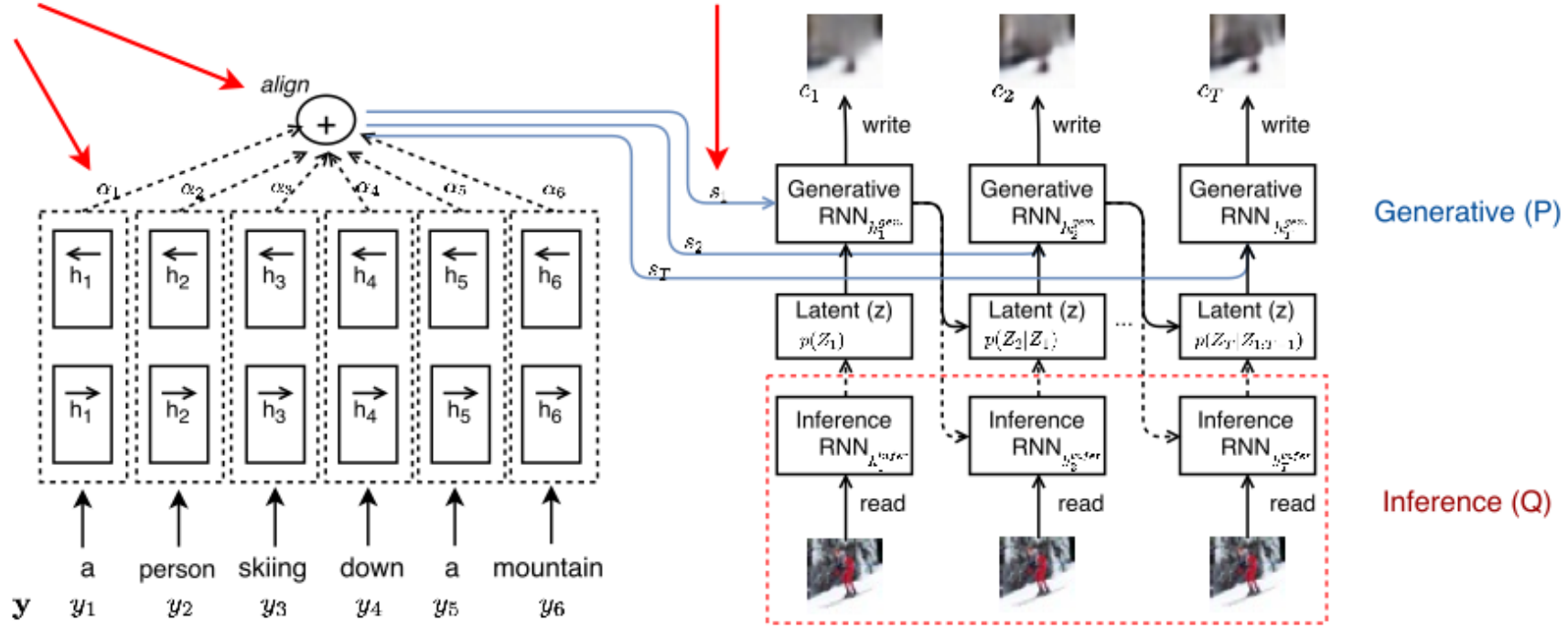
$$\mathbf{z}_t \sim P(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}) = \mathcal{N}\big(\mu(\mathbf{h}_{t-1}^{gen}), \sigma(\mathbf{h}_{t-1}^{gen})\big), \quad P(\mathbf{Z}_1) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

$$s_t = align\big(\mathbf{h}_{t-1}^{gen}, \mathbf{h}^{lang}\big) \quad \mathbf{h}_t^{gen} = LSTM^{gen}\big(\mathbf{h}_{t-1}^{gen}, [\mathbf{z}_t, s_t]\big)$$

$$\mathbf{c}_t = \mathbf{c}_{t-1} + write(\mathbf{h}_t^{gen}) \quad \mathbf{x} \sim P(\mathbf{x}|\mathbf{y}, \mathbf{Z}_{1:T}) = \prod_i \mathrm{Bern}(\boldsymbol{\sigma}(c_{T,i}))$$

- In practice, we use the conditional mean: $\mathbf{x} = \boldsymbol{\sigma}(\mathbf{c}_T)$.

# Alignment Model



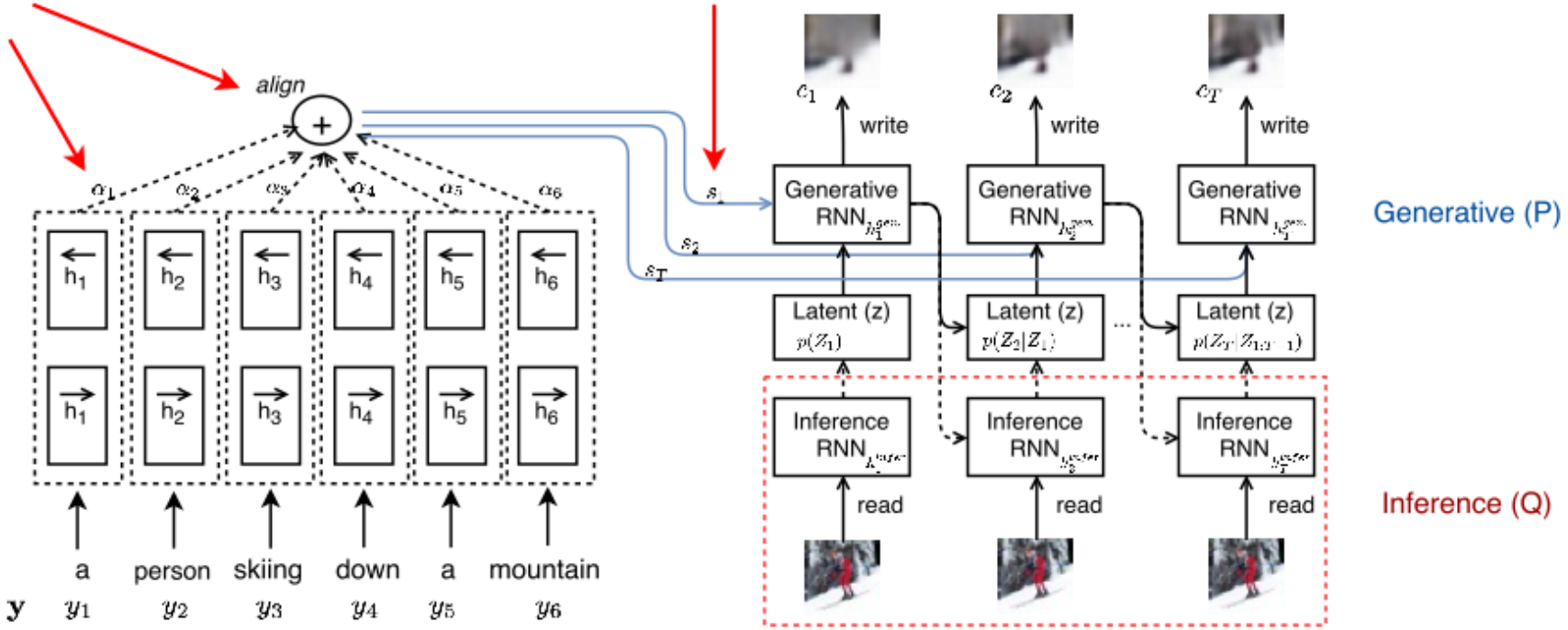

• Dynamic sentence representation at time t: weighted average of the bi-directional hidden states:

$$s_t = align\left(\mathbf{h}_{t-1}^{gen}, \mathbf{h}^{lang}\right) = \alpha_1^t \mathbf{h}_1^{lang} + \alpha_2^t \mathbf{h}_2^{lang} + \cdots + \alpha_N^t \mathbf{h}_N^{lang}$$

where the alignment probabilities are computed as:

$$e_k^t = \mathbf{v}^\top \tanh\left(U\mathbf{h}_k^{lang} + W\mathbf{h}_{t-1}^{gen} + b\right), \;\; \alpha_k^t = \frac{\exp\left(e_k^t\right)}{\sum_{i=1}^{N} \exp\left(e_i^t\right)}$$

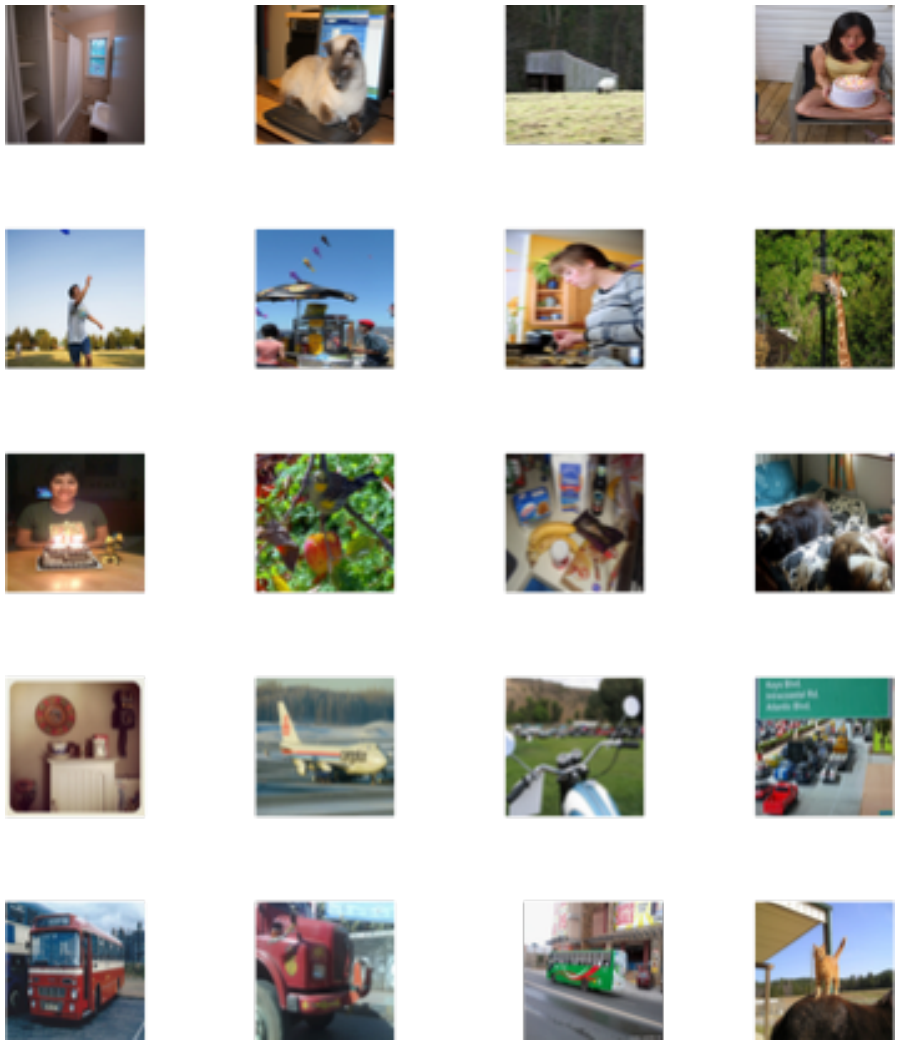
# Learning




• Maximize the variational lower bound on the marginal log-likelihood of the correct image **x** given the caption **y**:

$$\mathcal{L} = \sum_{Z} Q(Z|\mathbf{x},\mathbf{y}) \log P(\mathbf{x}|Z,\mathbf{y}) - \mathrm{D}_{KL}\big(Q(Z|\mathbf{x},\mathbf{y})||P(Z|\mathbf{y})\big)$$
$$\leq \log P(\mathbf{x}|\mathbf{y})$$

# MS COCO Dataset



- Contains 83K images.
- Each image contains 5 captions.
- Standard benchmark dataset for many of the recent image captioning systems.

Lin et. al. 2014

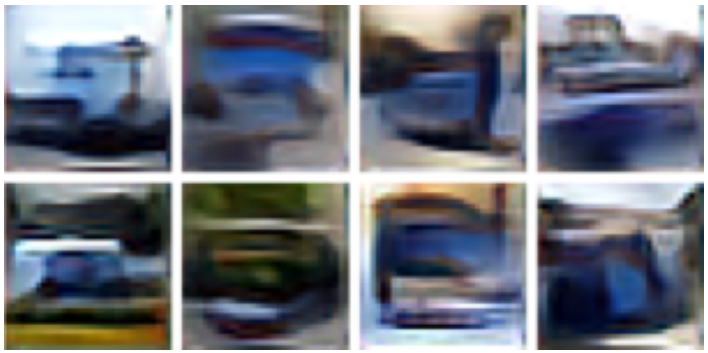
# Flipping Colors

A **yellow school bus** parked in the parking lot

A **red school bus** parked in the parking lot

A **green school bus** parked in the parking lot

A **blue school bus** parked in the parking lot



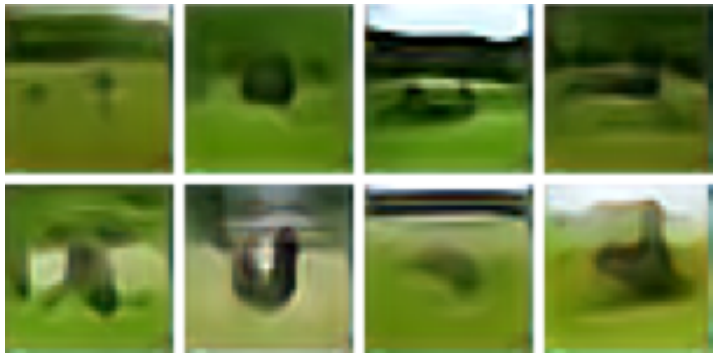
# Flipping Backgrounds

A very large commercial plane flying **in clear skies**.

A very large commercial plane flying **in rainy skies**.

A herd of elephants walking across a **dry grass field**.

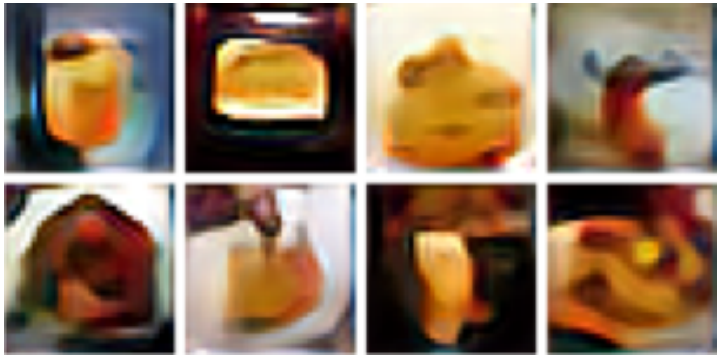
A herd of elephants walking across a **green grass field**.



# Flipping Objects

**The decadent chocolate desert** is on the table.

**A bowl of bananas** is on the table..



A vintage photo of **a cat**.

A vintage photo of **a dog**.

# Qualitative Comparison

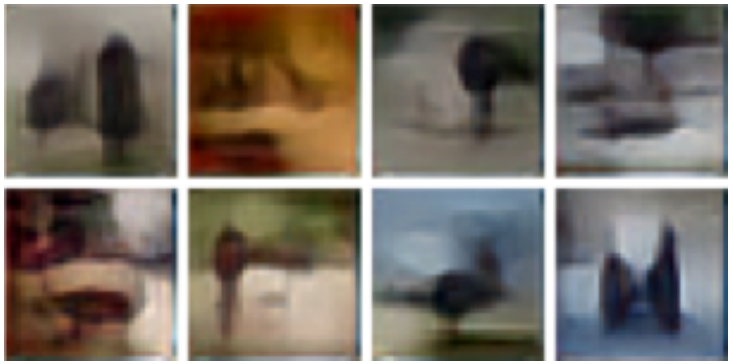
*A group of people walk on a beach with surf boards*

Our Model

LAPGAN (Denton et. al. 2015)

Conv-Deconv VAE

Fully Connected VAE



# Variational Lower-Bound

• We can estimate the variational lower-bound on the average test log-probabilities:

| Model | Training | Test |
|---|---|---|
| Our Model | -1792,15 | -1791,53 |
| Skipthought-Draw | -1794,29 | -1791,37 |
| noAlignDraw | -1792,14 | -1791,15 |

• At least we can see that we do not overfit to the training data, unlike many other approaches.

# Novel Scene Compositions

A toilet seat sits open in the bathroom

A toilet seat sits open in the grass field

Ask Google?

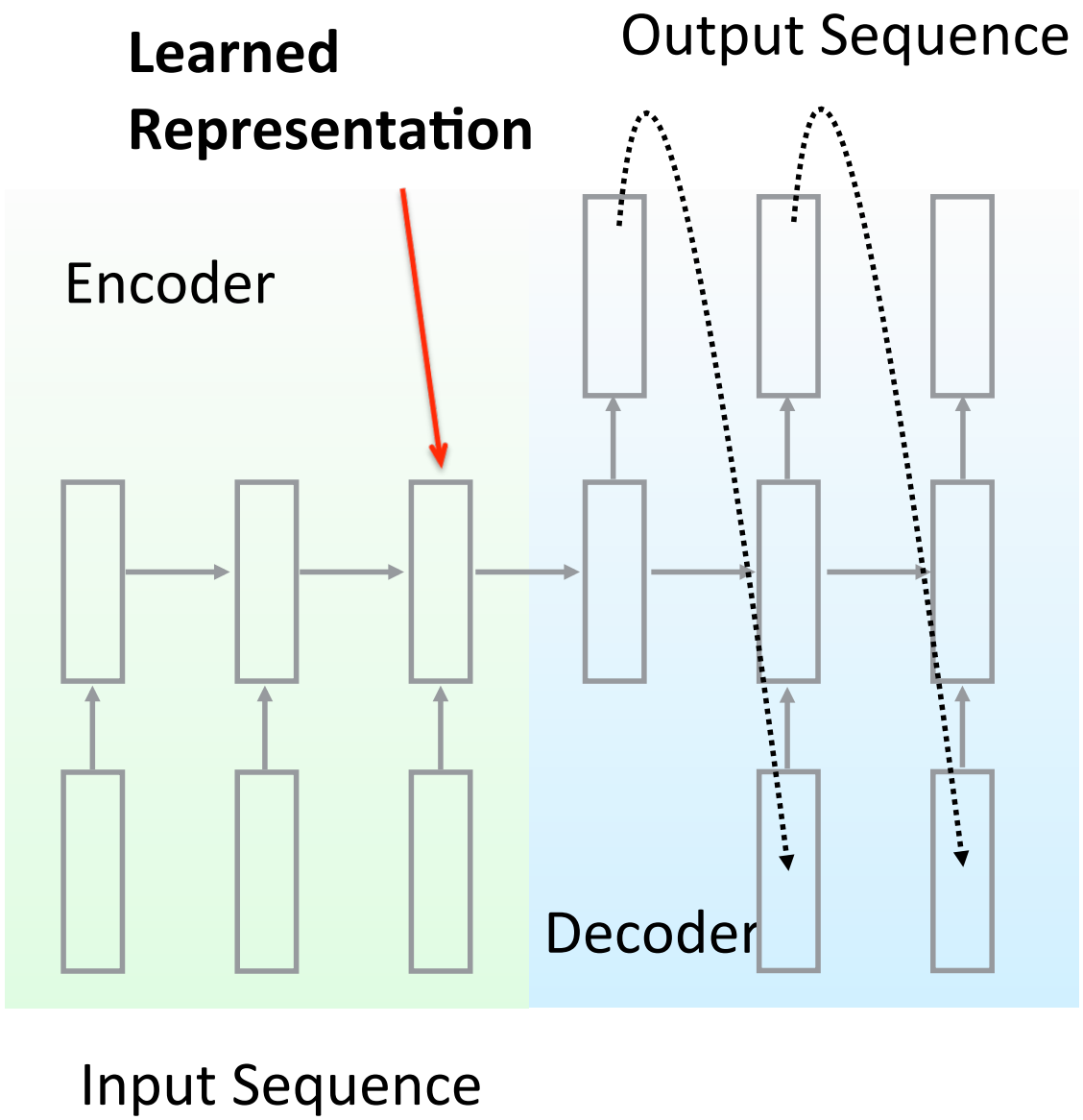
# (Some) Open Problems

- Unsupervised Learning / Transfer Learning / One-Shot Learning
- Reasoning, Attention, and Memory
- Natural Language Understanding
- Deep Reinforcement Learning

# (Some) Open Problems

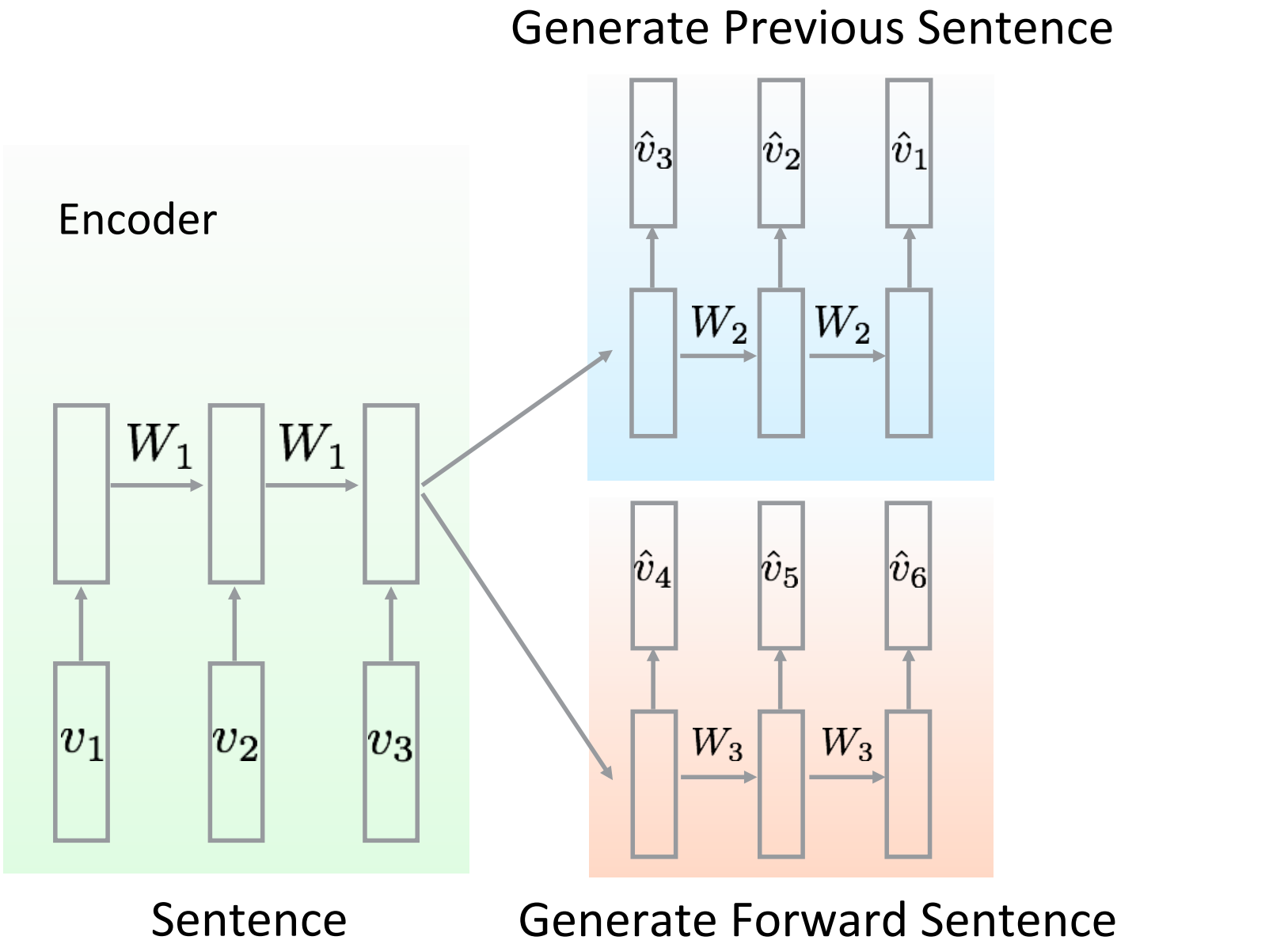
- Unsupervised Learning / Transfer Learning / One-Shot Learning
- Reasoning, Attention, and Memory
- Natural Language Understanding
- Deep Reinforcement Learning

# Sequence to Sequence Learning




- RNN Encoder-Decoders for Machine Translation (Sutskever et al. 2014; Cho et al. 2014; Kalchbrenner et al. 2013, Srivastava et al., 2015)

# Skip-Thought Model




(Kiros et al., NIPS 2015)

# Learning Objective

• Objective: The sum of the log-probabilities for the next and previous sentences conditioned on the encoder representation:

representation of encoder

$$\sum_t \log P(w_{i+1}^t | w_{i+1}^{<t}, \mathbf{h}_i) + \sum_t \log P(w_{i-1}^t | w_{i-1}^{<t}, \mathbf{h}_i)$$

Forward sentence

Previous sentence

• Data: Book-11K corpus:

| # of books | # of sentences | # of words | # of unique words |
|---|---|---|---|
| 11,038 | 74,004,228 | 984,846,357 | 1,316,420 |

# Semantic Relatedness

| | Method | $r$ | $\rho$ | MSE |
|---|---|---|---|---|
| SemEval 2014 sub-missions | Illinois-LH [18] | 0.7993 | 0.7538 | 0.3692 |
| | UNAL-NLP [19] | 0.8070 | 0.7489 | 0.3550 |
| | Meaning Factory [20] | 0.8268 | 0.7721 | 0.3224 |
| | ECNU [21] | 0.8414 | – | – |
| Results reported by Tai et.al. | Mean vectors [22] | 0.7577 | 0.6738 | 0.4557 |
| | DT-RNN [23] | 0.7923 | 0.7319 | 0.3822 |
| | SDT-RNN [23] | 0.7900 | 0.7304 | 0.3848 |
| | LSTM [22] | 0.8528 | 0.7911 | 0.2831 |
| | Bidirectional LSTM [22] | 0.8567 | 0.7966 | 0.2736 |
| | Dependency Tree-LSTM [22] | **0.8676** | **0.8083** | **0.2532** |
| Ours | uni-skip | 0.8477 | 0.7780 | 0.2872 |
| | bi-skip | 0.8405 | 0.7696 | 0.2995 |
| | combine-skip | 0.8584 | 0.7916 | 0.2687 |
| | combine-skip+COCO | 0.8655 | 0.7995 | 0.2561 |

- Outperform all previous systems from the SemEval 2014 competition.

# Semantic Relatedness Recurrent Neural Network

- How similar the two sentences are on the scale 1 to 5?

Ground Truth 5.0 Prediction 4.9

| | |
|---|---|
| A man is driving a car. | A car is being driven by a man. |

Ground Truth 2.9 Prediction 3.5

| | |
|---|---|
| A girl is looking at a woman in costume. | A girl in costume looks like a woman. |

Ground Truth 2.6 Prediction 4.4

| | |
|---|---|
| A person is performing tricks on a motorcycle | The performer is tricking a person on a motorcycle |

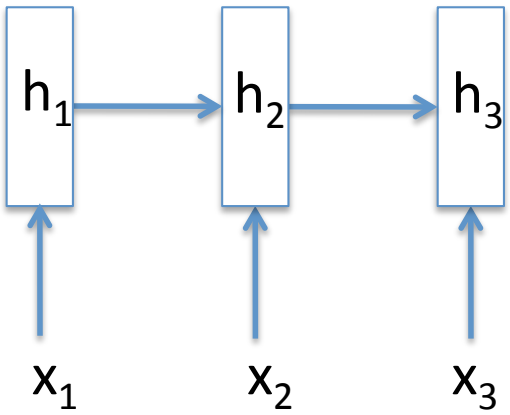
# Recurrent Neural Network

$$\mathbf{h_t} = \phi\left(\mathbf{U}\mathbf{h_{t-1}} + \mathbf{W}\mathbf{x_t} + \mathbf{b}\right)$$

Nonlinearity

Hidden State at previous time step

Input at time step t



# Multiplicative Integration

- Replace

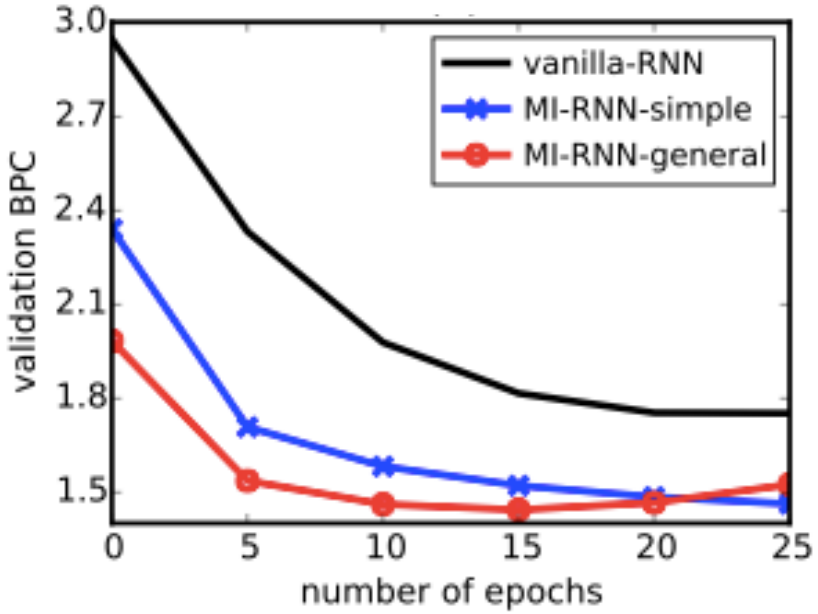
$$\phi(\mathbf{Uh} + \mathbf{Wx} + \mathbf{b})$$

- With

$$\phi(\mathbf{Uh} \odot \mathbf{Wx} + \mathbf{b})$$

- Or more generally

$$\phi(\alpha \odot \mathbf{Uh} \odot \mathbf{Wx} + \beta_1 \odot \mathbf{Uh} + \beta_2 \odot \mathbf{Wx} + \mathbf{b})$$




(Wu et al., NIPS 2016)

# Who-Did-What Dataset

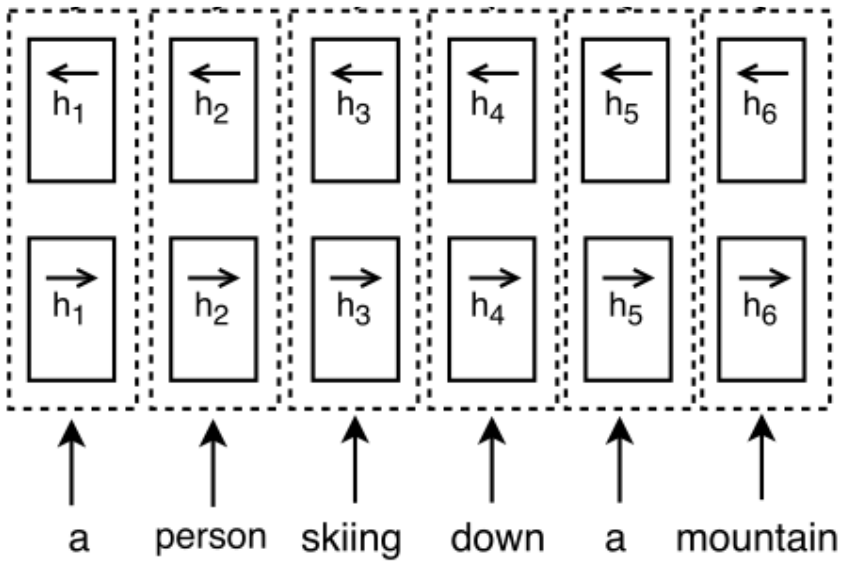
- **Document**: "…arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges … included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama…"

- **Query**: President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat.

- **Answer**: Rod Blagojevich

Onishi, Wang, Bansal, Gimpel, McAllester, EMNLP, 2016

# Representing Document/Query




- Forward RNN reads sentences from left to right:

$$[\overrightarrow{h}_1, \overrightarrow{h}_2, .., \overrightarrow{h}_{|D|}]$$

- Backward RNN reads sentences from right to left:

$$[\overleftarrow{h}_1, \overleftarrow{h}_2, .., \overleftarrow{h}_{|D|}]$$

- The hidden states are then concatenated:

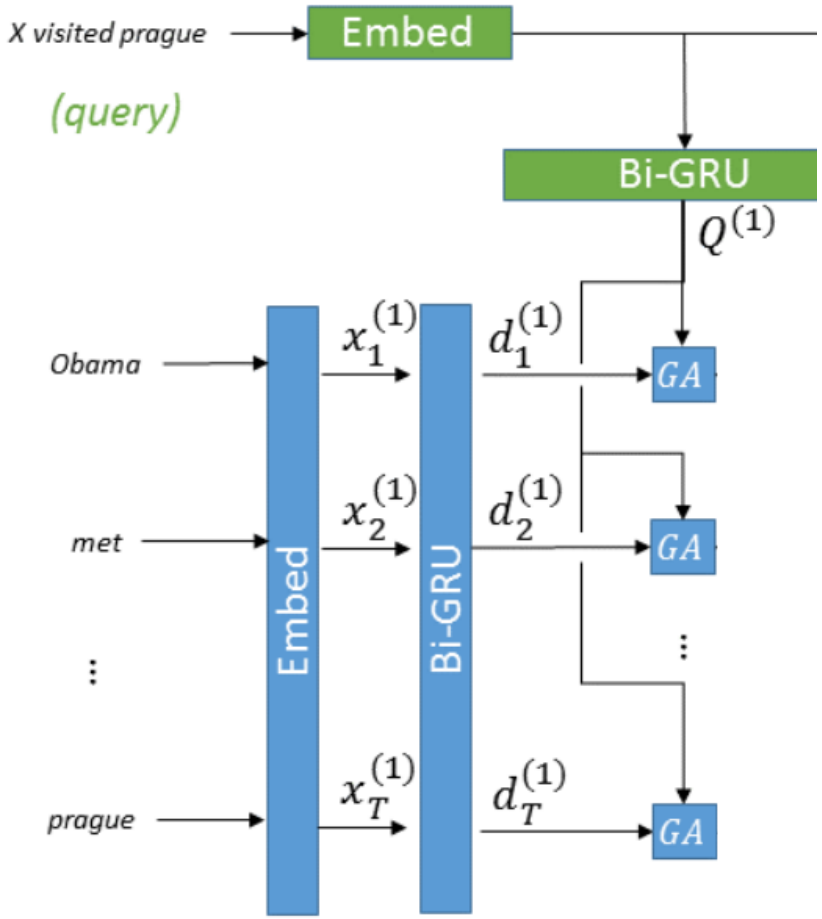
$$\overleftrightarrow{\text{GRU}} = [h_1, h_2, ..., h_{|D|}], \quad h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i]$$

- Use GRUs to encode a document and a query:

$$D = \overleftrightarrow{\text{GRU}}_D(X) \quad Q = \overleftrightarrow{\text{GRU}}_Q(Y)$$

# Gated Attention (GA) Mechanism

• For each word in document D, we form a token-specific representation of the query Q:




$$\alpha_i = \text{softmax}(Q^\top d_i)$$

$$\tilde{q}_i = Q\alpha_i$$

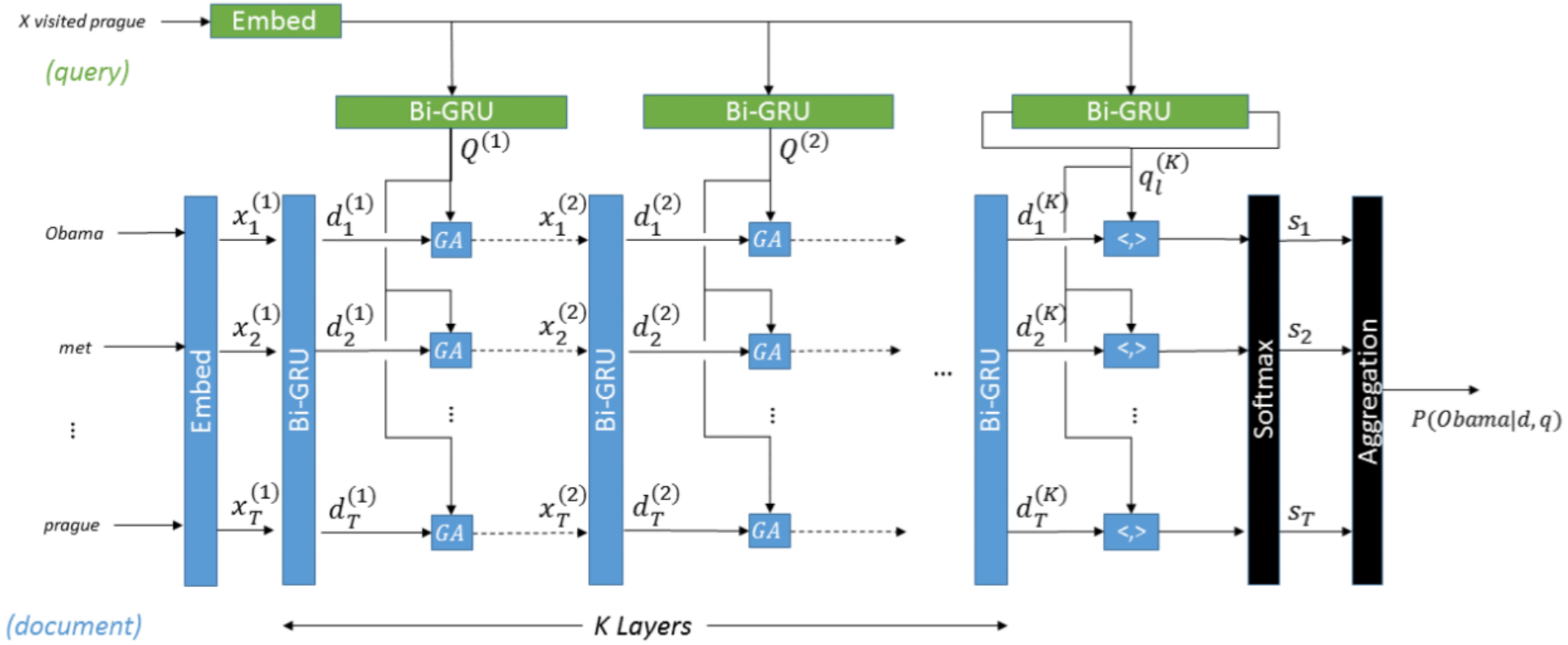
$$x_i = d_i \odot \tilde{q}_i$$

➢ use the element-wise multiplication operator to model the interactions between $d_i$ and $\tilde{q}_i$

(Dhingra, Liu, Yang, Cohen, Salakhutdinov, 2016)

# Multi-hop Architecture

- Reasoning over multiple sentences requires several passes over the context



# Affect of Multiplicative Gating

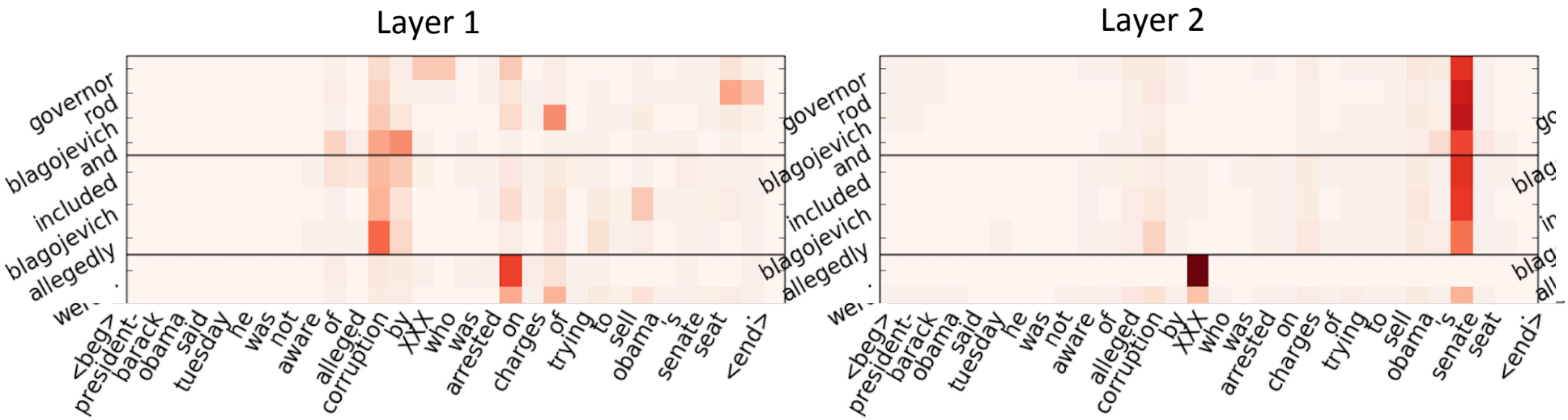
• Performance of different gating functions on “Who did What” (WDW) dataset.

| Gating Function | Accuracy | |
|---|---|---|
| | Val | Test |
| Sum | 62.9 | 62.1 |
| Concatenate | 63.1 | 61.1 |
| Multiply | **67.8** | **67.0** |

| Model | Strict | | Relaxed | |
|---|---|---|---|---|
| | Val | Test | Val | Test |
| Human † | – | 84.0 | – | – |
| Attentive Reader † | – | 53.0 | – | 55.0 |
| AS Reader † | – | 57.0 | – | 59.0 |
| Stanford AR † | – | 64.0 | – | 65.0 |
| NSE † | 66.5 | 66.2 | 67.0 | 66.7 |
| GA Reader-- † | – | 57.0 | – | 60.0 |
| GA Reader | 67.8 | 67.0 | 66.4 | 66.3 |
| GA Reader (+feature) | **70.1** | **69.5** | **70.9** | **70.6** |

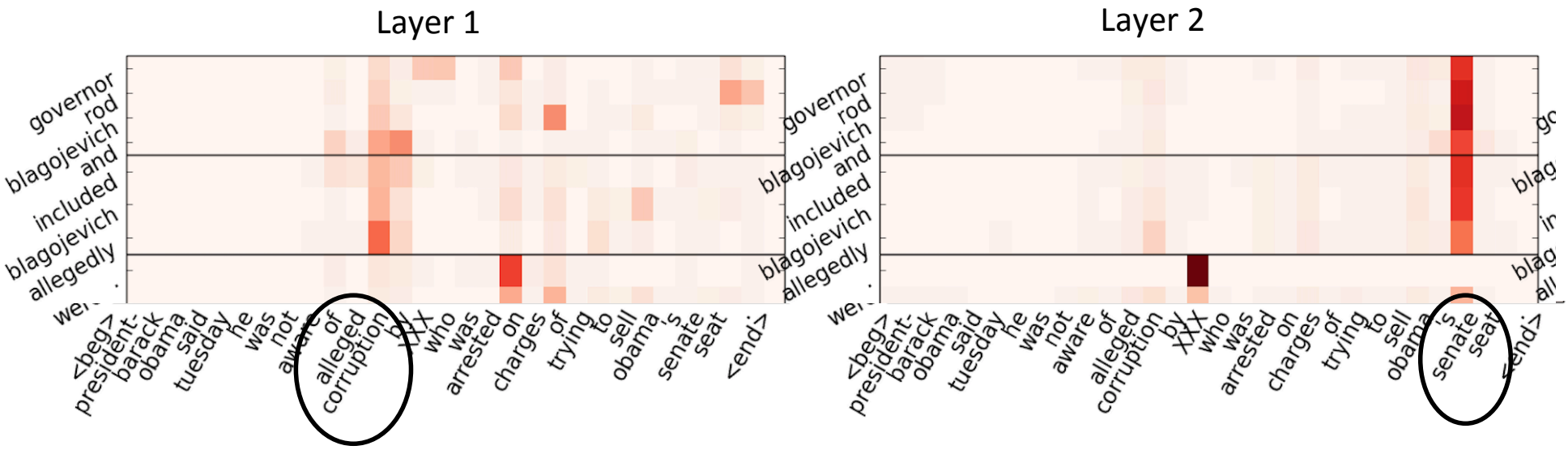
# Analysis of Attention

- **Context**: "…arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges … included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama…"
- **Query**: "President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
- **Answer**: **Rod Blagojevich**



# Analysis of Attention

- **Context**: "…arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges … included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama…"
- **Query**: "President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
- **Answer**: **Rod Blagojevich**




Code + Data: https://github.com/bdhingra/ga-reader

# Broad-Context Language Modeling

Her plain face broke into a huge smile when she saw Terry.
“Terry!” she called out.
She rushed to meet him and they embraced.
“Hon, I want you to meet an old friend, Owen McKenna.
Owen, please meet Emily."
She gave me a quick nod and turned back to **X**

LAMBADA dataset, Paperno et al., 2016

# Broad-Context Language Modeling

Her plain face broke into a huge smile when she saw **Terry**.
“Terry!” she called out.
She rushed to meet him and they embraced.
“Hon, I want you to meet an old friend, **Owen** McKenna.
Owen, please meet **Emily**."
She gave me a quick nod and turned back to **X**

LAMBADA dataset, Paperno et al., 2016

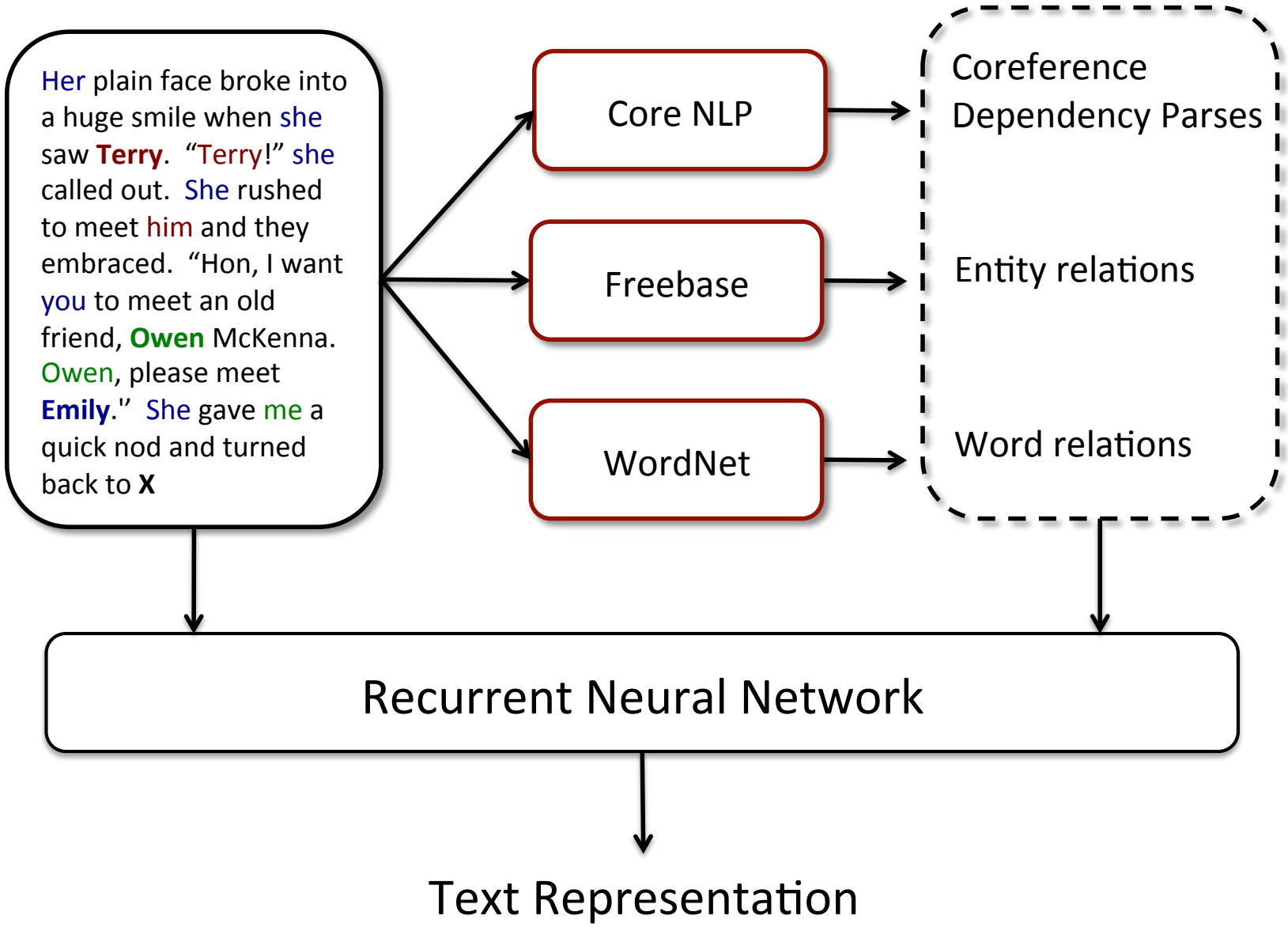
# Broad-Context Language Modeling

Her plain face broke into a huge smile when she saw **Terry**.
"Terry!" she called out.
She rushed to meet him and they embraced.
"Hon, I want you to meet an old friend, **Owen** McKenna.
Owen, please meet **Emily**."
She gave me a quick nod and turned back to **X**

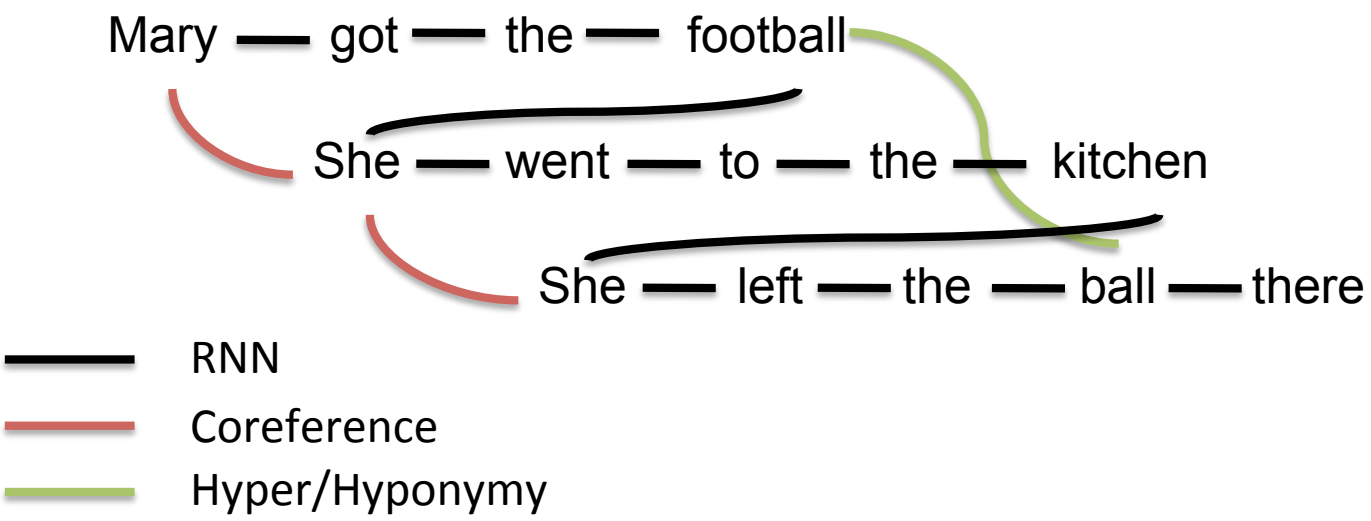
X = **Terry**

LAMBADA dataset, Paperno et al., 2016

# Incorporating Prior Knowledge

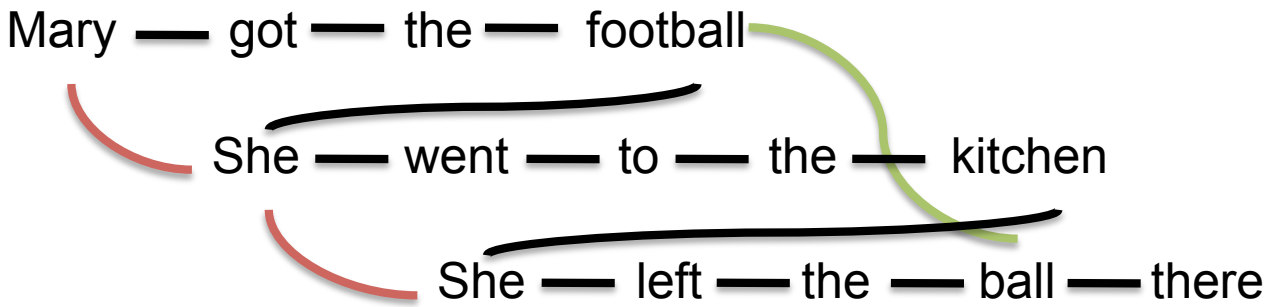


# Incorporating Prior Knowledge




Dhingra, Yang, Cohen, Salakhutdinov 2017

# Incorporating Prior Knowledge

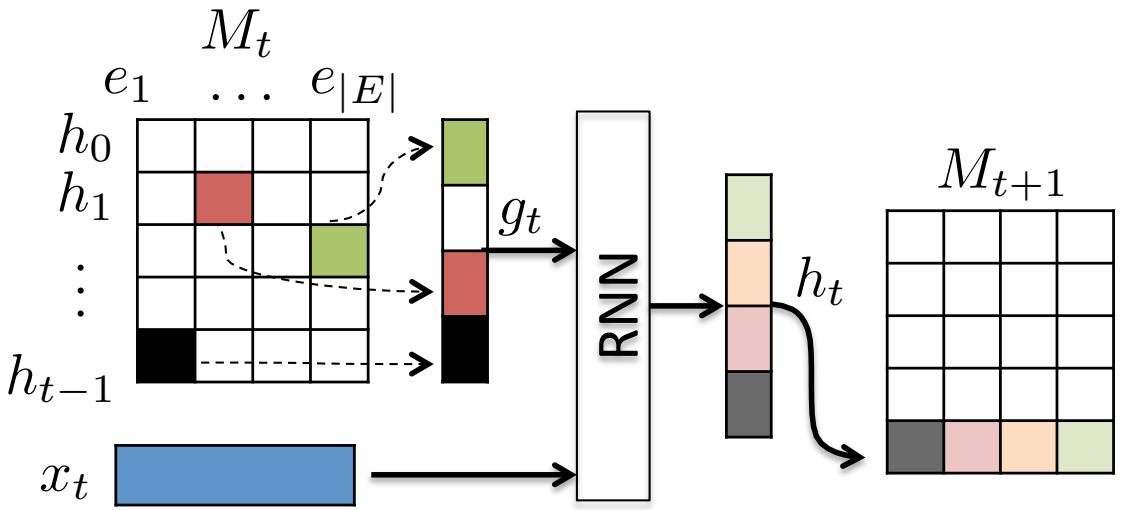



RNN

Coreference

Hyper/Hyponymy

**M**emory as **A**cyclic **G**raph **E**ncoding (MAGE) - RNN

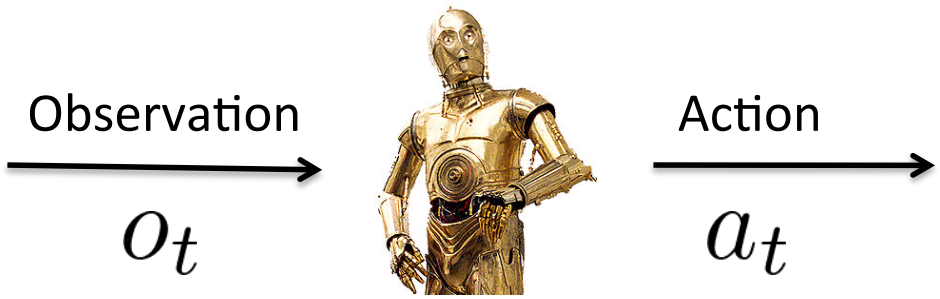



Dhingra, Yang, Cohen, Salakhutdinov 2017

# (Some) Open Problems

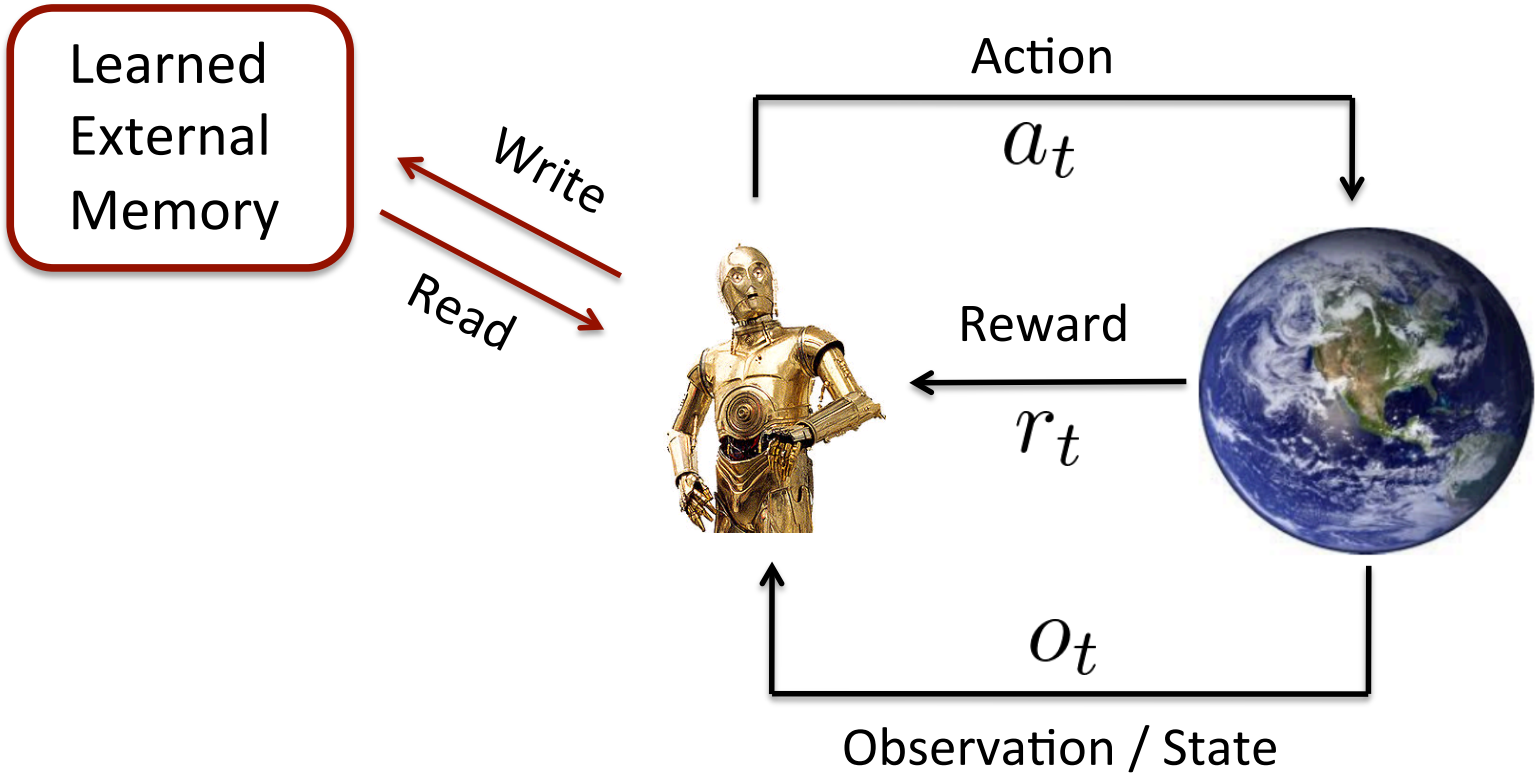
- Unsupervised Learning / Transfer Learning / One-Shot Learning
- Reasoning, Attention, and Memory
- Natural Language Understanding
- Deep Reinforcement Learning

# Learning Behaviors



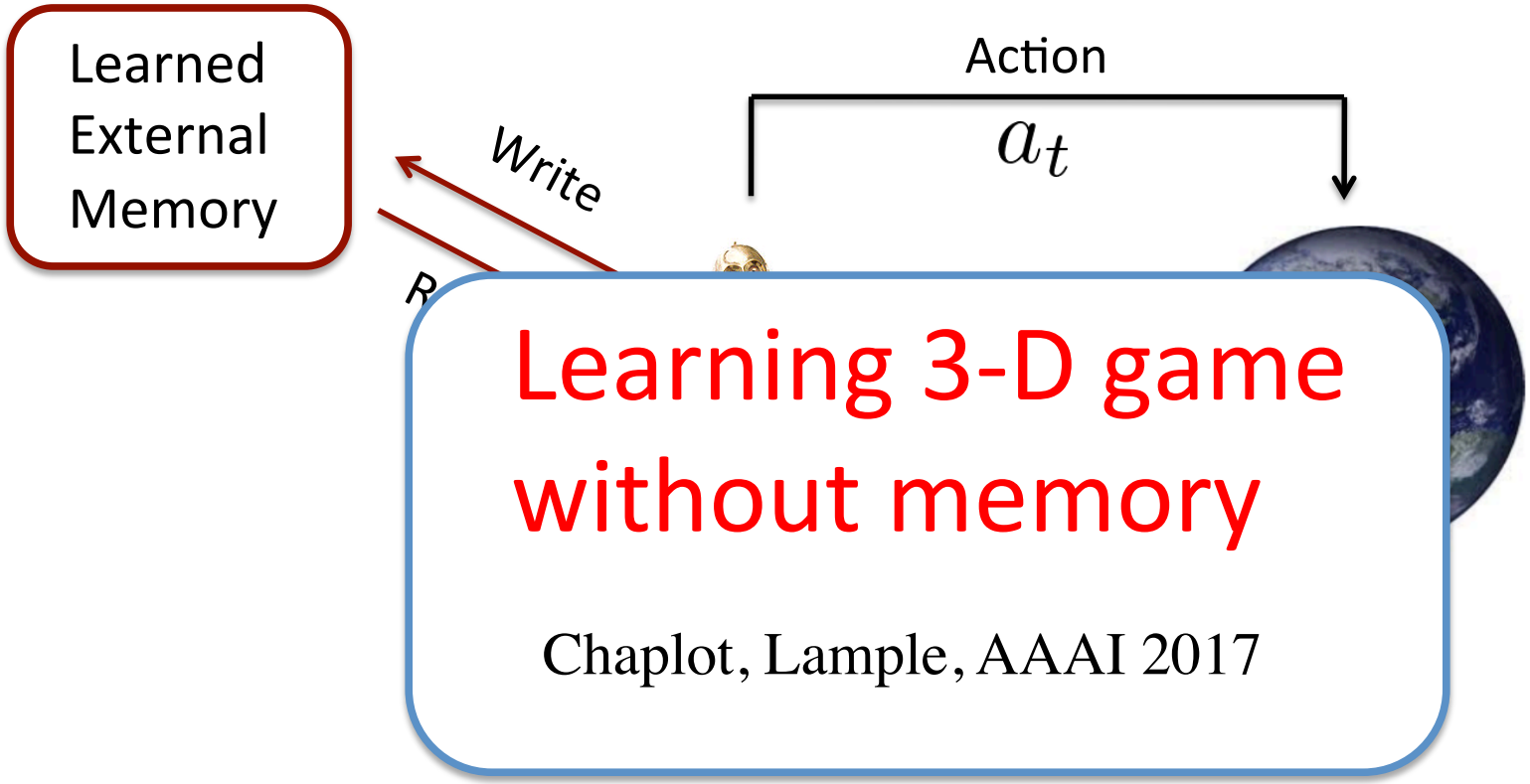

Learning to map sequences of observations to actions, for a particular goal

# Reinforcement Learning with Memory



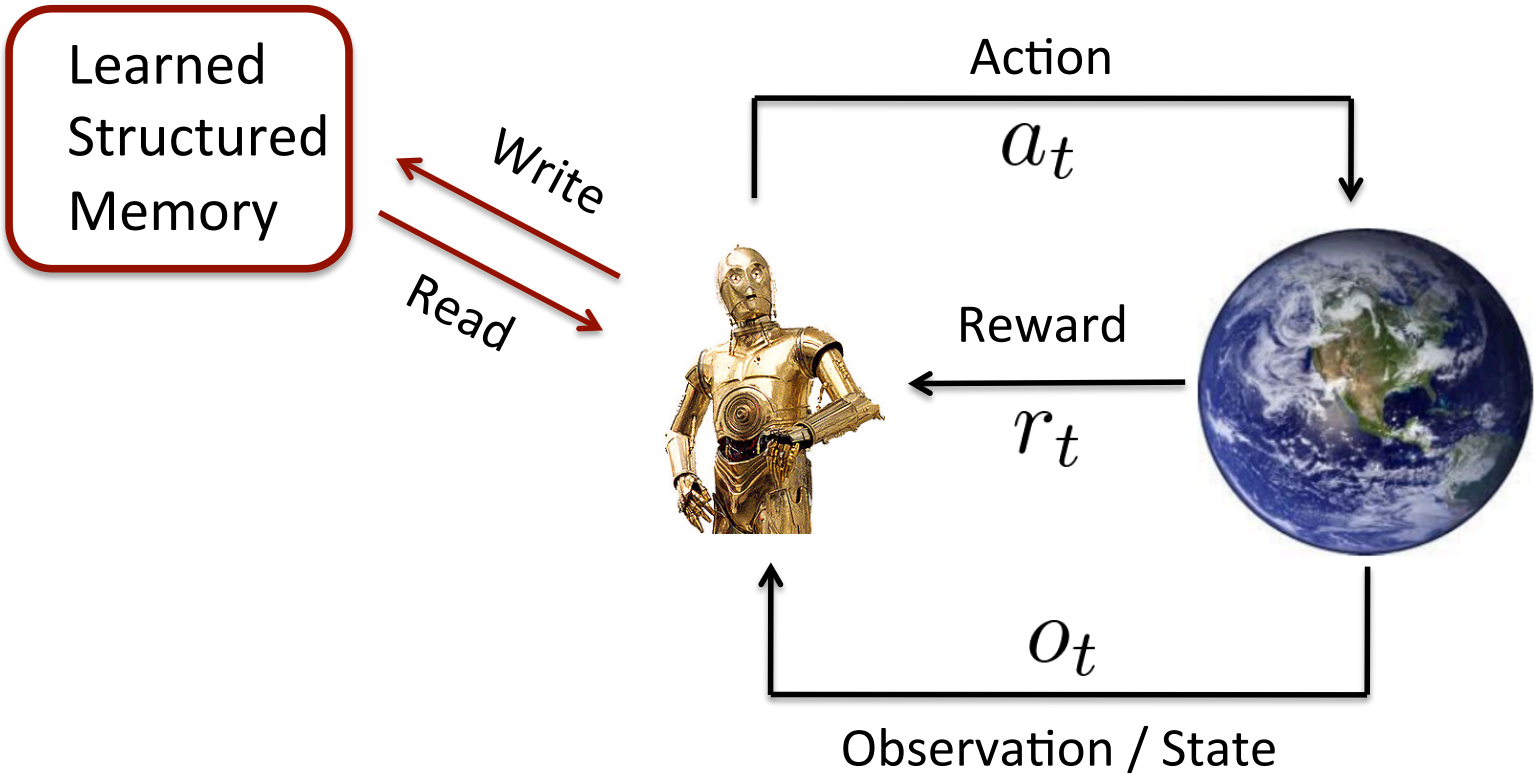

Differentiable Neural Computer, Graves et al., Nature, 2016;
Neural Turing Machine, Graves et al., 2014

# Reinforcement Learning with Memory




Differentiable Neural Computer, Graves et al., Nature, 2016;
Neural Turing Machine, Graves et al., 2014

# Deep RL with Memory

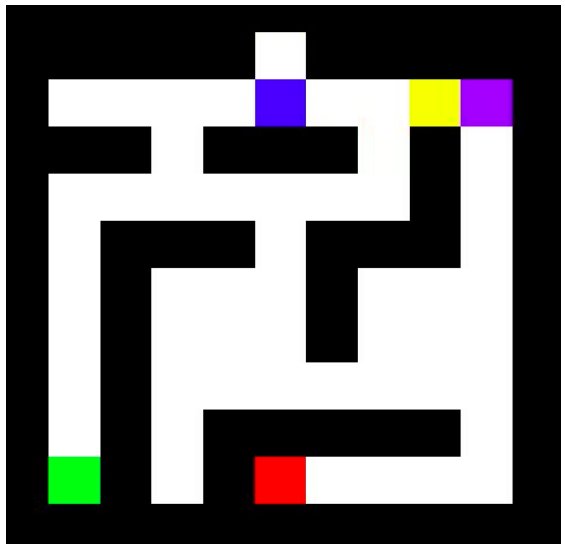



Parisotto, Salakhutdinov, 2017

# Random Maze with Indicator

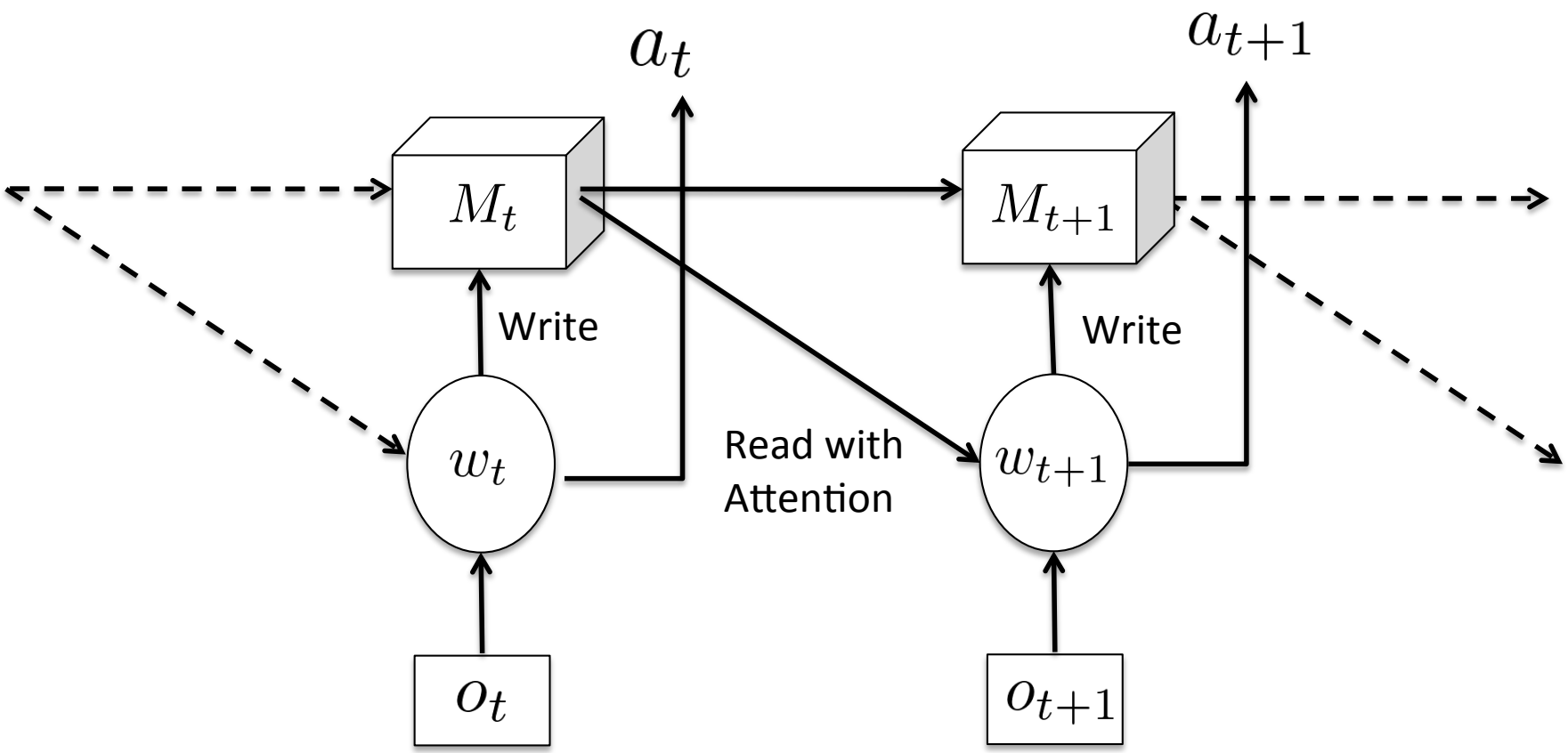
- Indicator: Either blue or pink
  - If blue, find the green block
  - If pink, find the red block

- Negative reward if agent does not find correct block in N steps or goes to wrong block.



Parisotto, Salakhutdinov, 2017

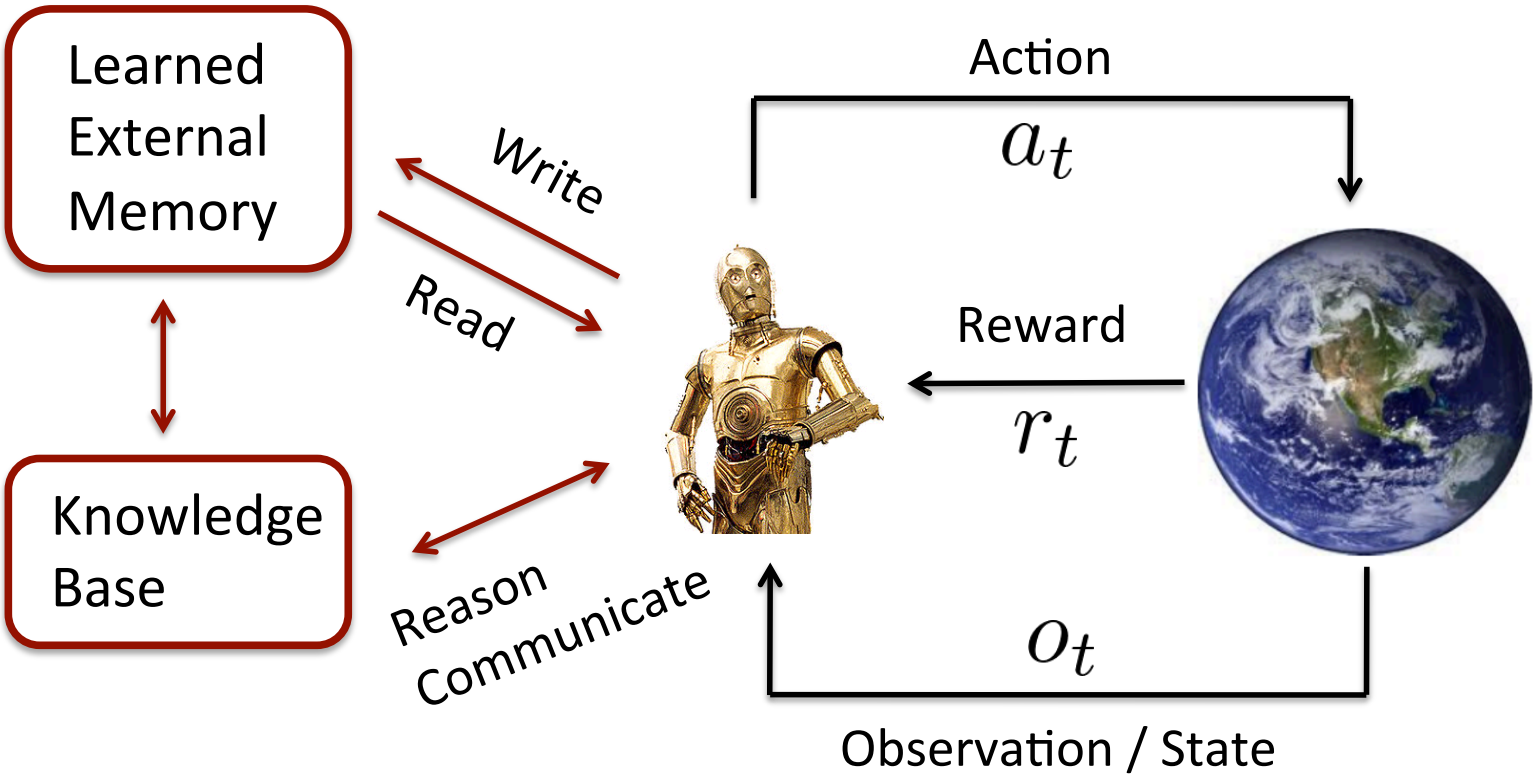
# Random Maze with Indicator



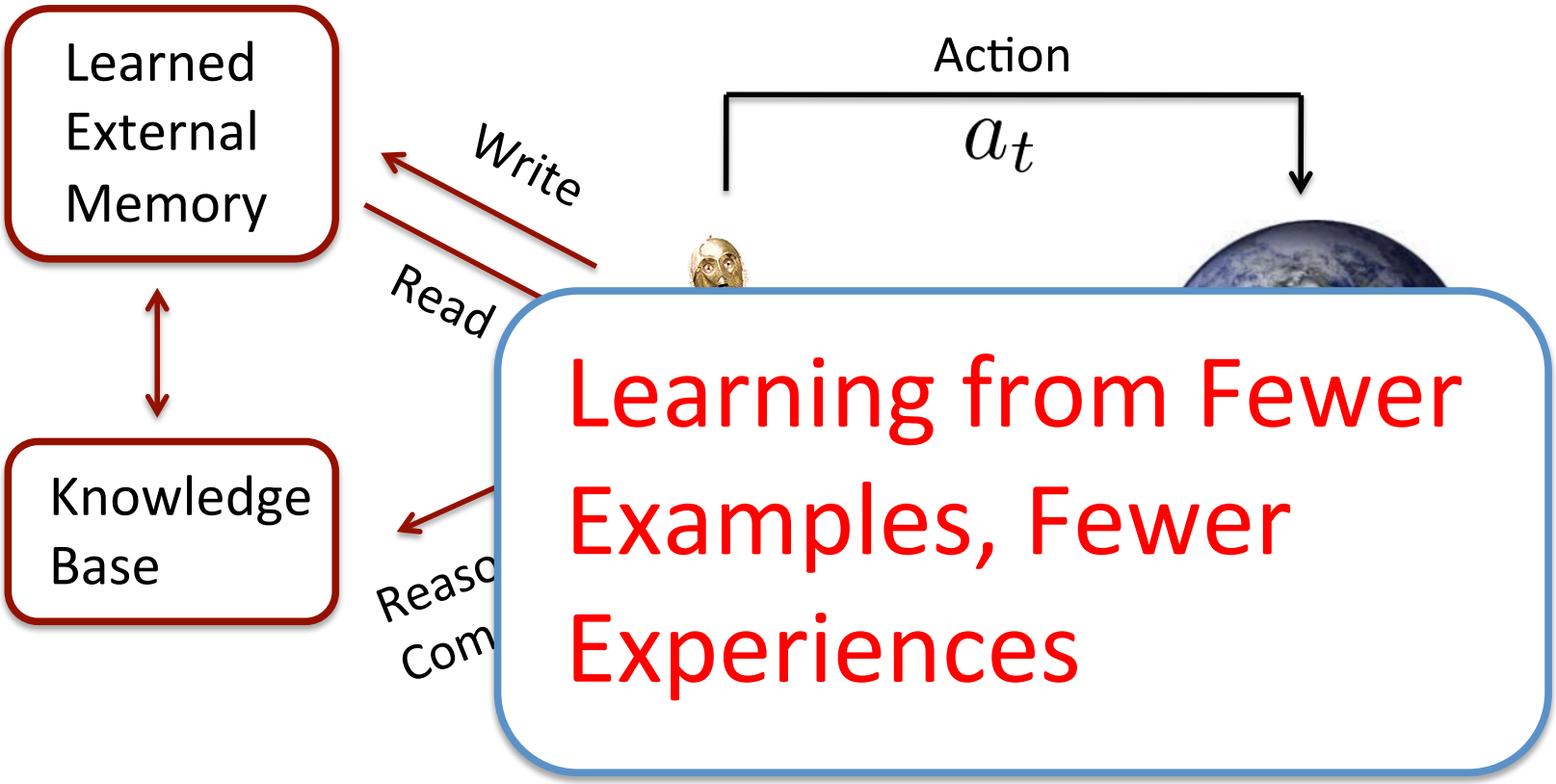

Parisotto, Salakhutdinov, 2017

# Random Maze with Indicator

# Building Intelligent Agents



# Building Intelligent Agents



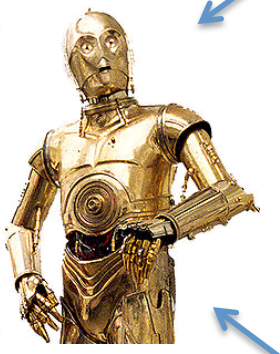
# Deep Multimodal Learning

Images

Text & Language

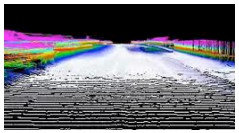

REUTERS

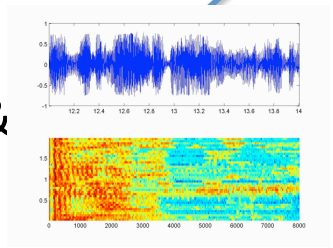
Associated Press



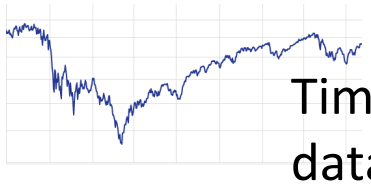
Video

Laser scans





Speech & Audio



Time series data

Develop learning systems that come closer to displaying human like intelligence

Thank you