

Beam Search for Top-B Decoding in Bi-RNNs

Qing Sun



Dhruv Batra



Deep Learning Summer School, Montreal, CA

Image captioning

[Karpathy *et al*, CVPR2015]

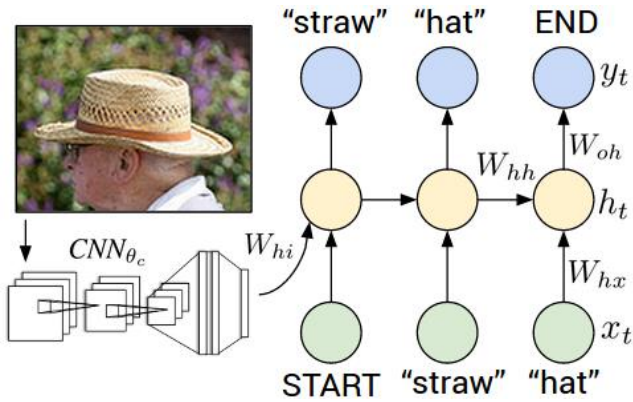
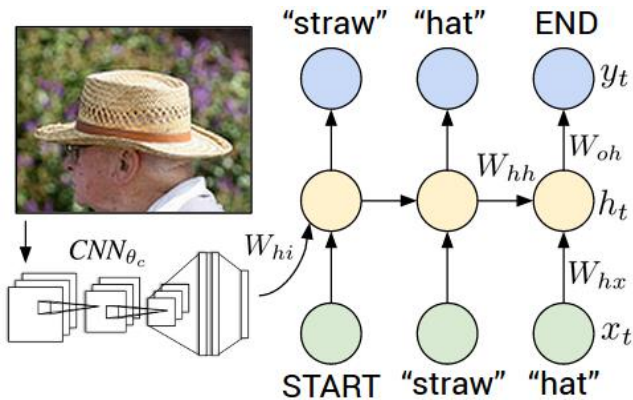


Image captioning

[Karpathy *et al*, CVPR2015]



Visual Question Answering

[Antol *et al*, ICCV 2015]

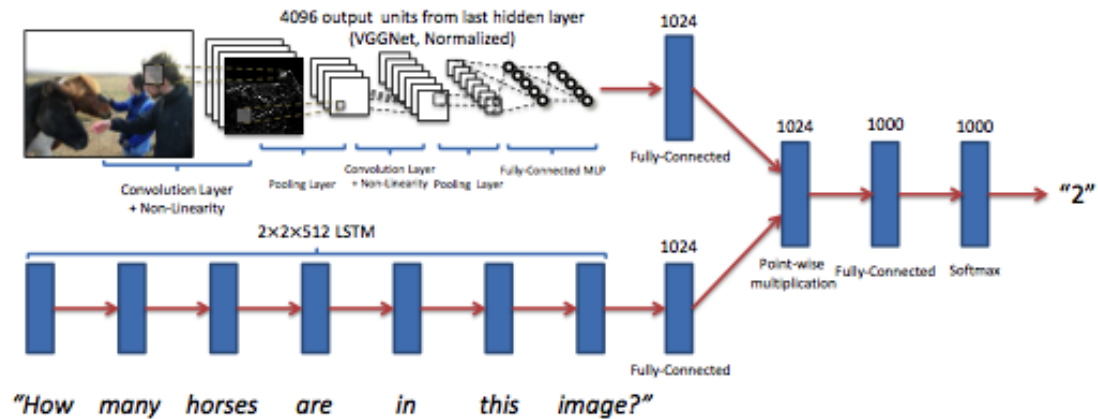
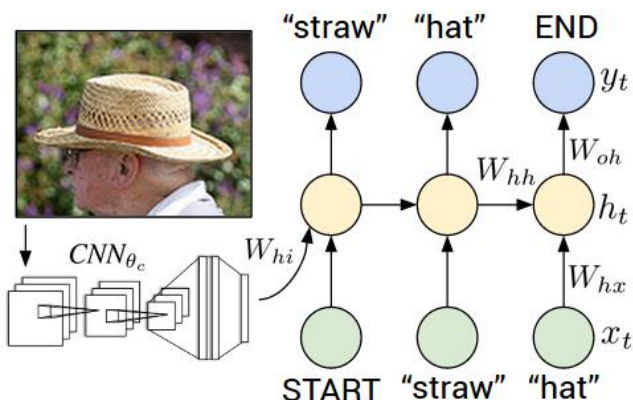


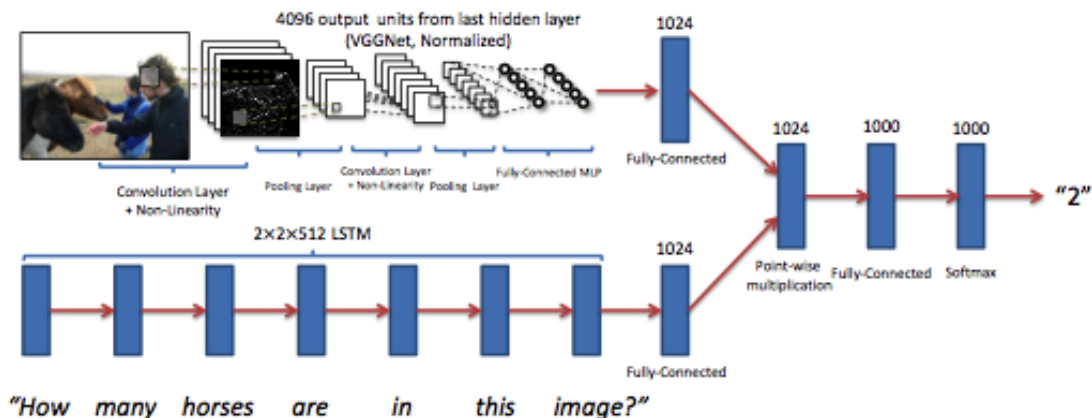
Image captioning

[Karpathy *et al*, CVPR2015]



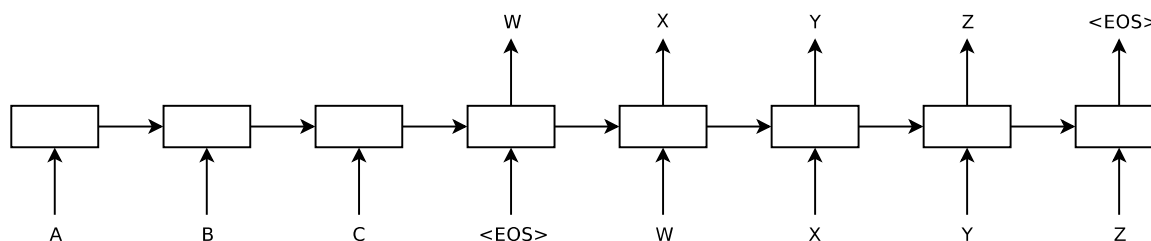
Visual Question Answering

[Antol *et al*, ICCV 2015]



Machine Translation

[Ilya Sutskever *et al*, NIPS 2014]



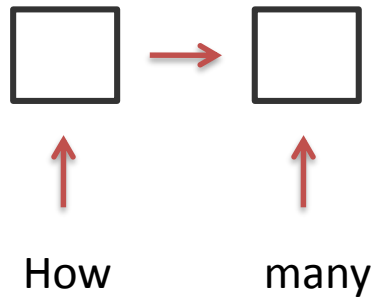
Encoder vs. Decoder

Encoder vs. Decoder

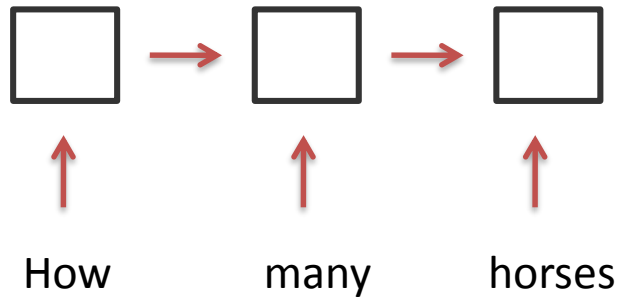


How

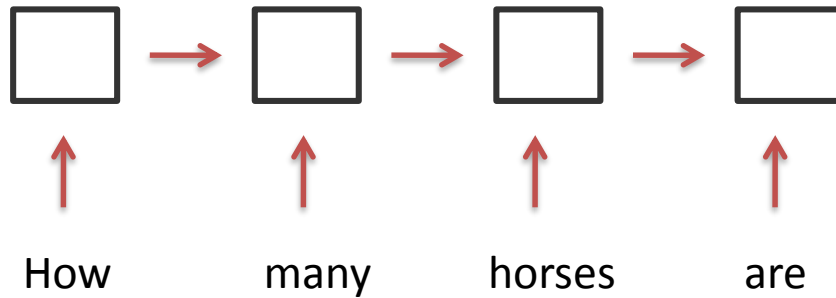
Encoder vs. Decoder



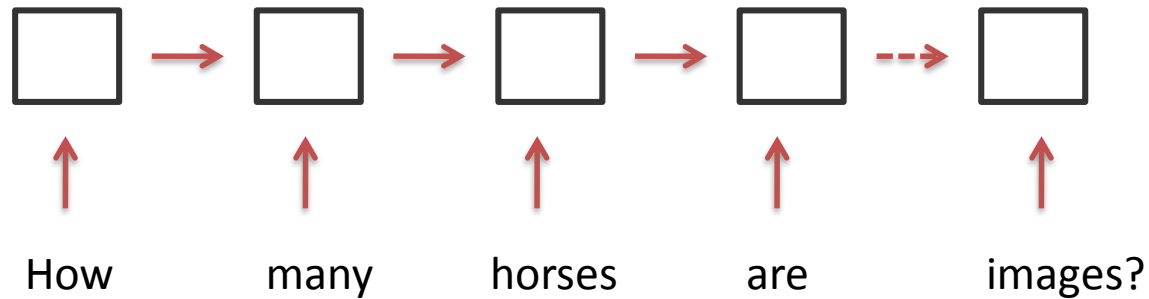
Encoder vs. Decoder



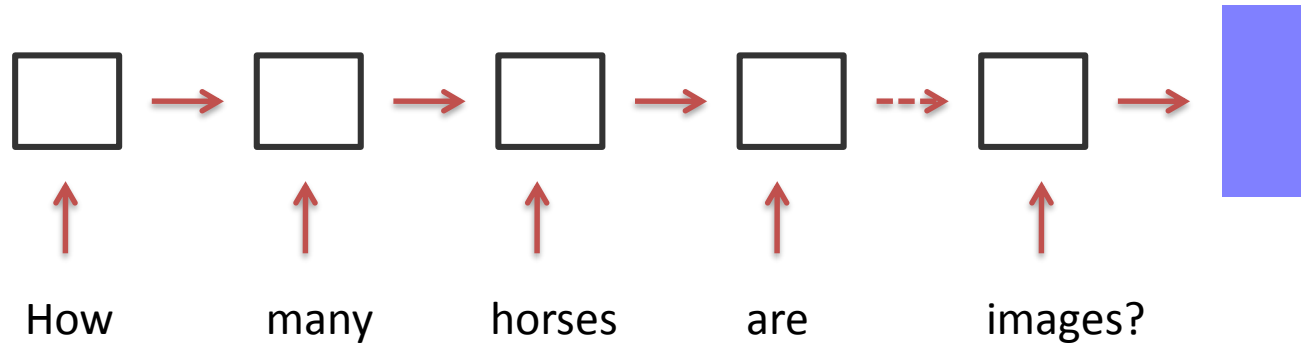
Encoder vs. Decoder



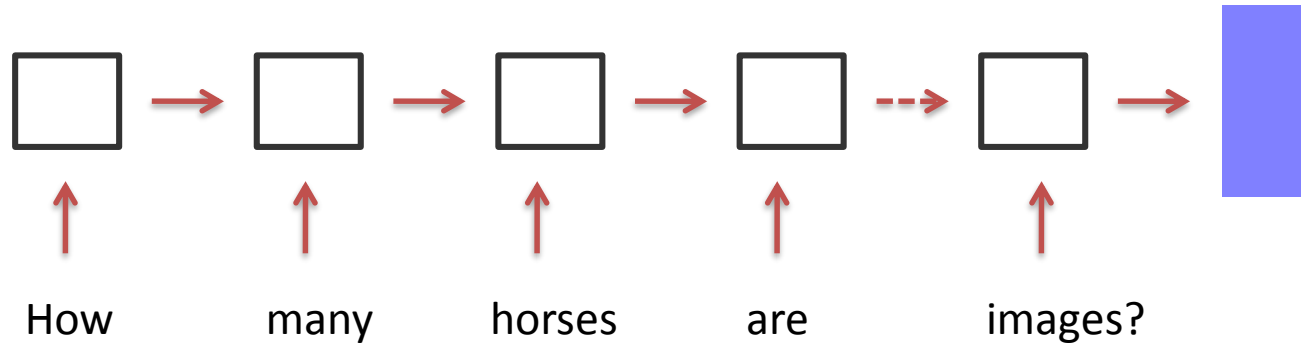
Encoder vs. Decoder



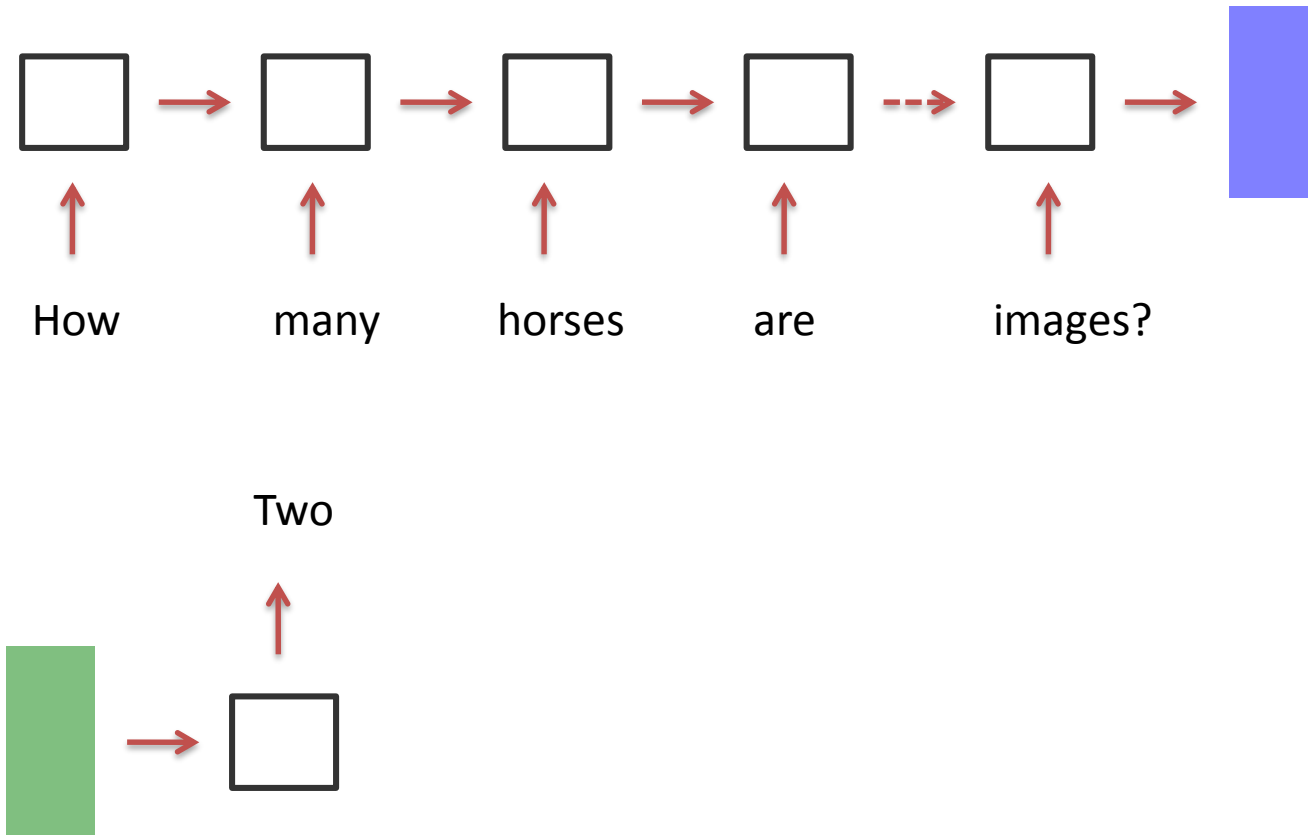
Encoder vs. Decoder



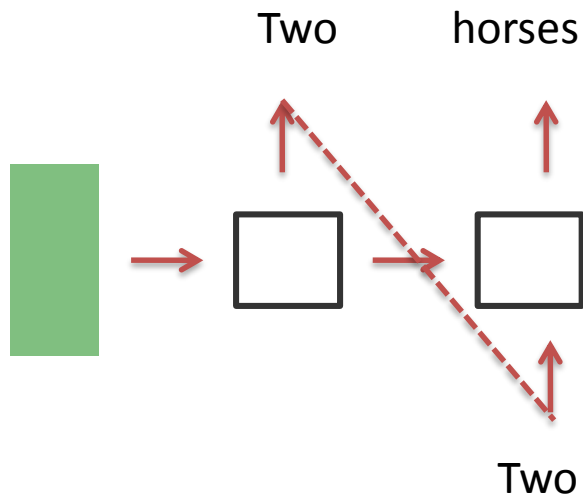
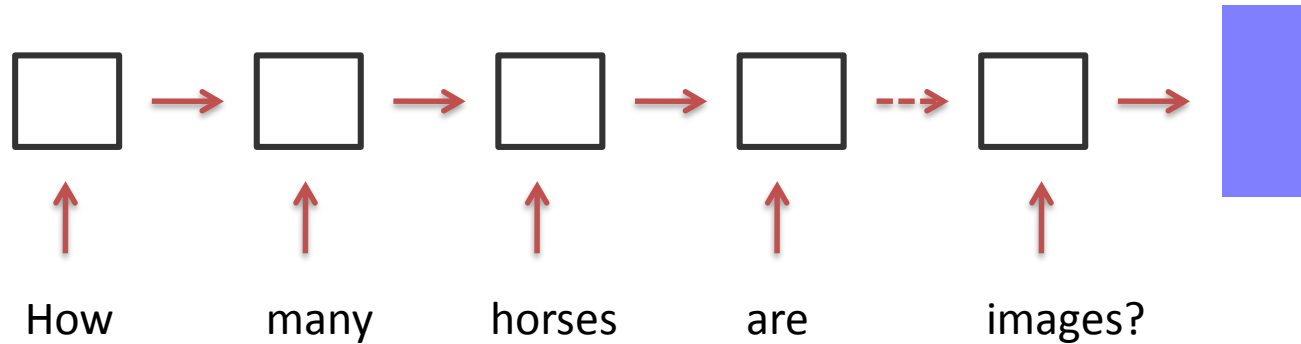
Encoder vs. Decoder



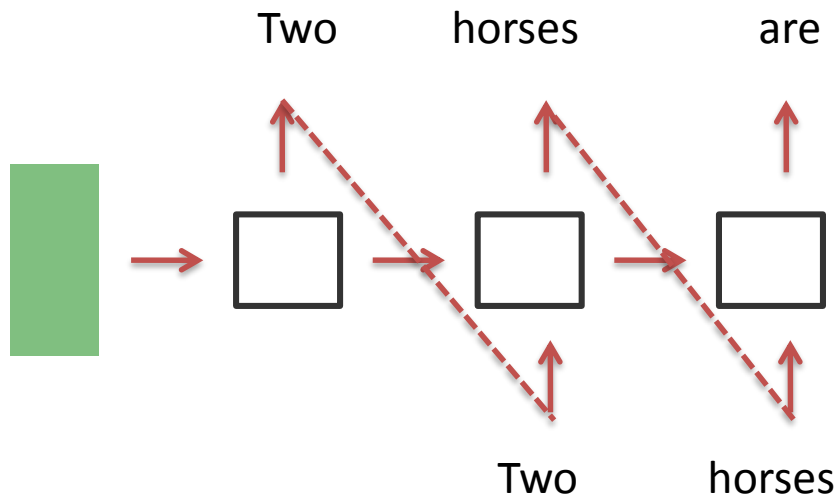
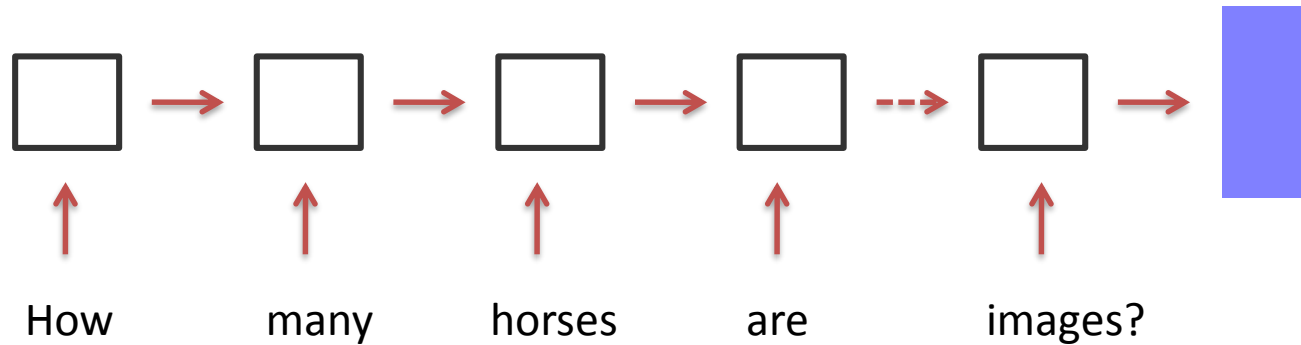
Encoder vs. Decoder



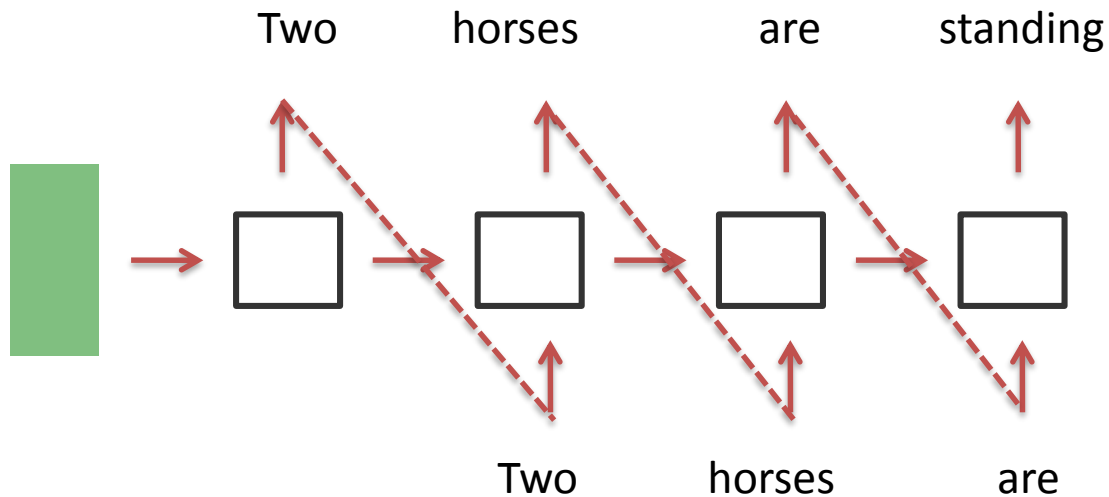
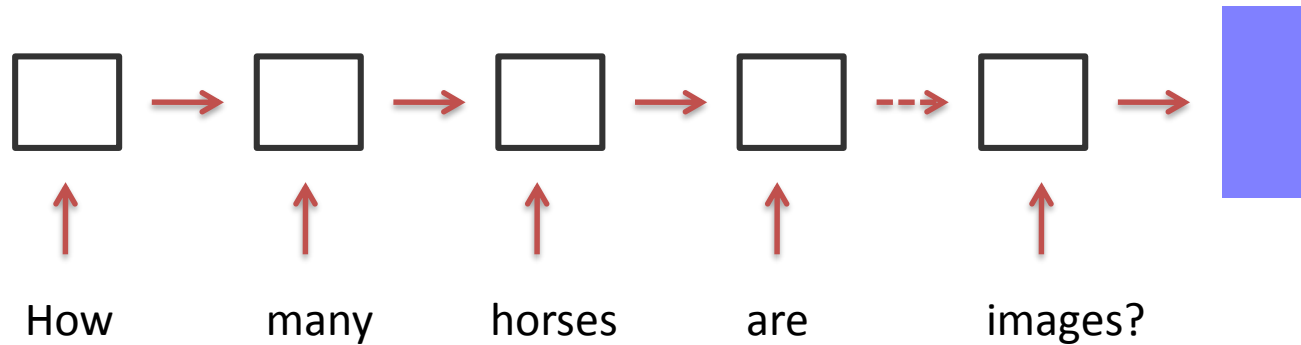
Encoder vs. Decoder



Encoder vs. Decoder



Encoder vs. Decoder



Encoder vs. Decoder

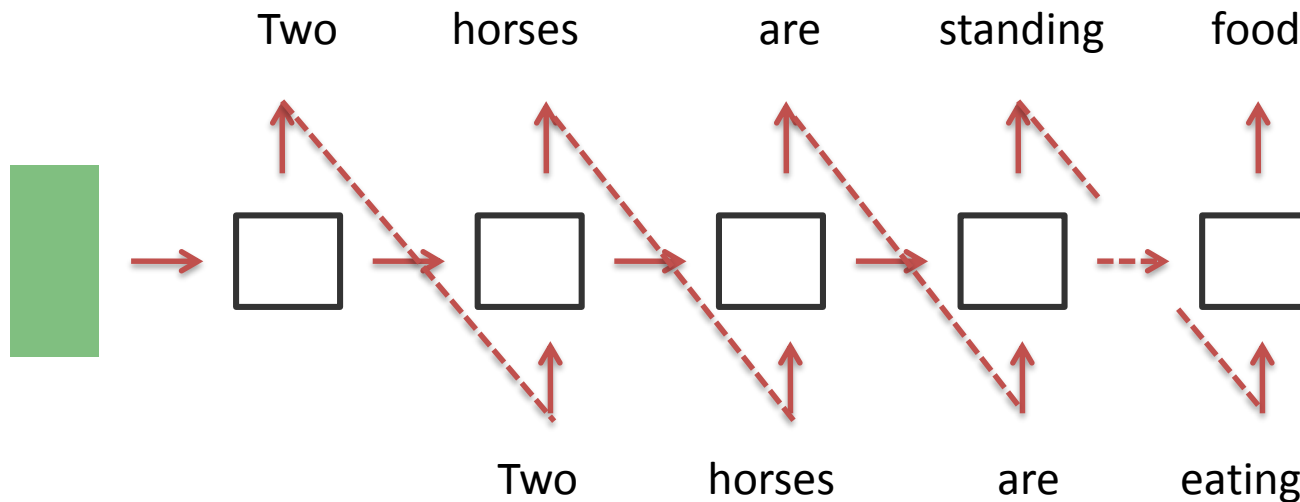
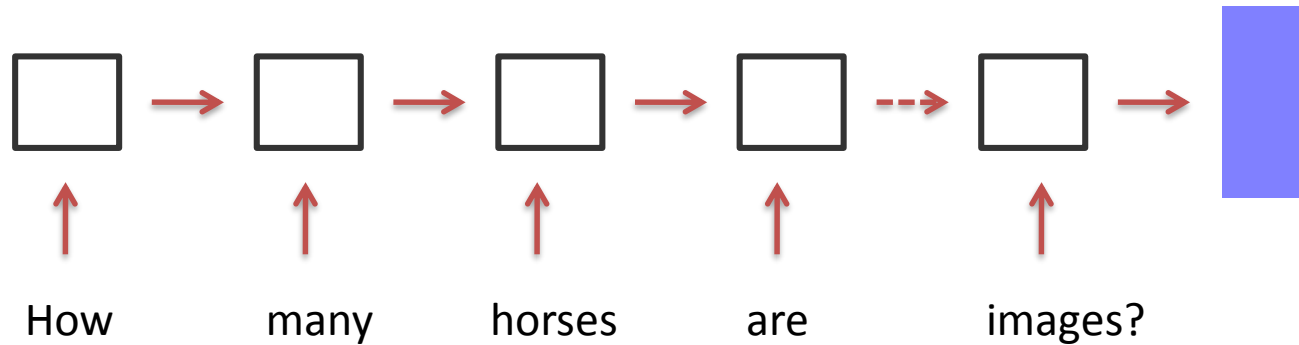
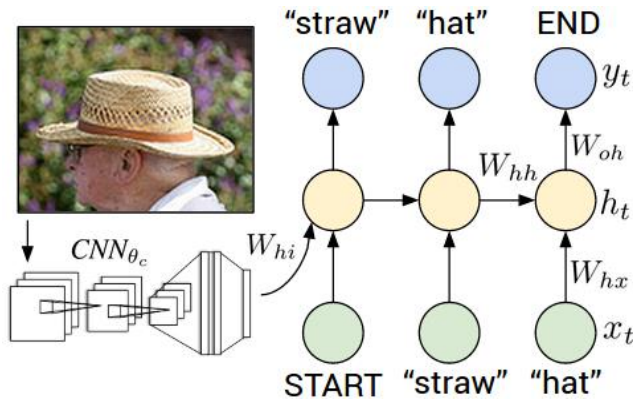


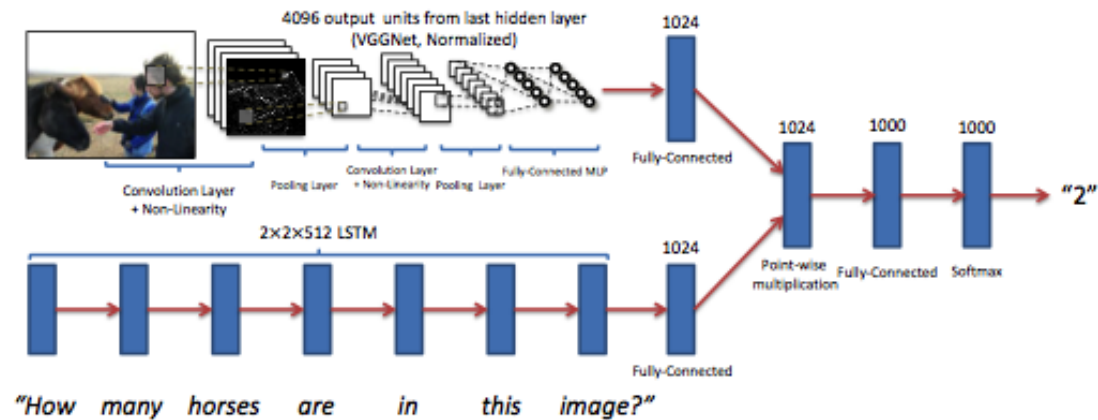
Image captioning

[Karpathy, *etal*, CVPR 2015]



Visual Question Answering

[Antol, *etal*, ICCV 2015]



Machine Translation

[Ilya Sutskever, *etal*, NIPS 2014]

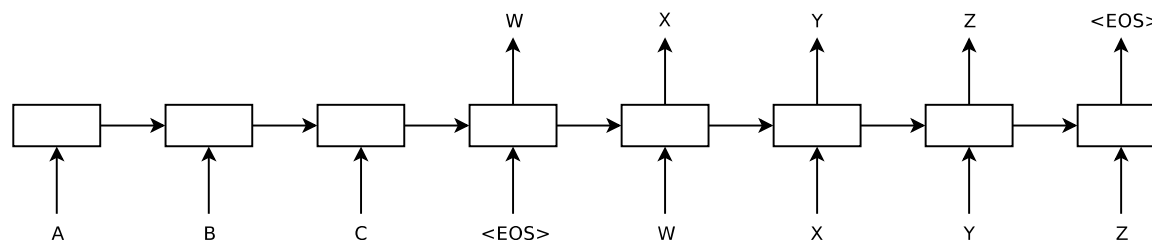
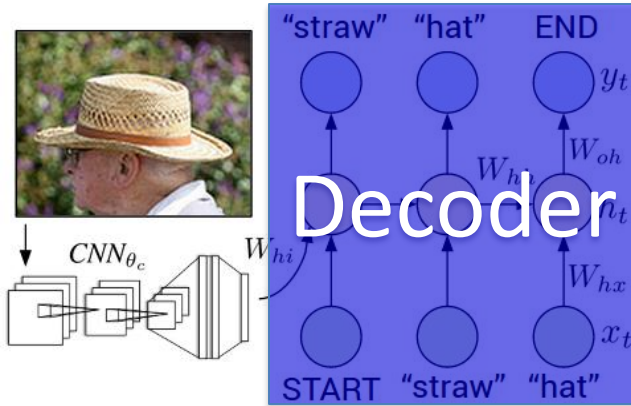


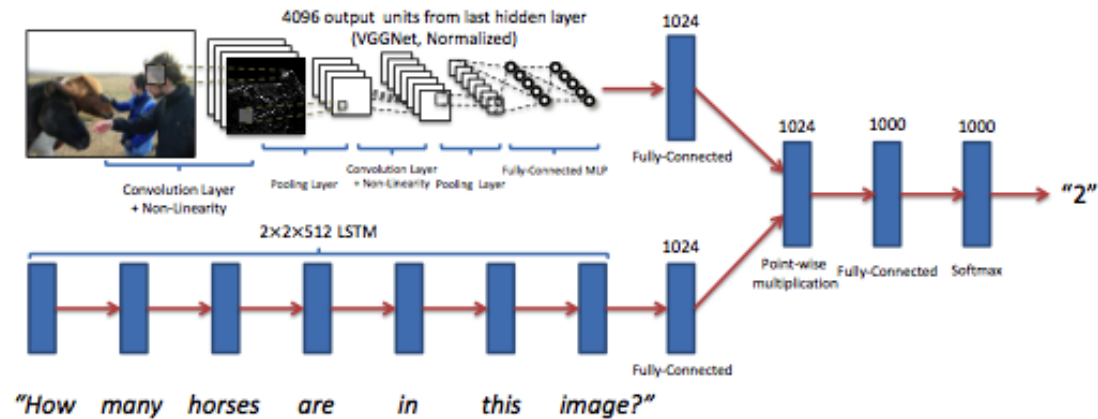
Image captioning

[Karpathy, *etal*, CVPR 2015]



Visual Question Answering

[Antol, *etal*, ICCV 2015]



Machine Translation

[Ilya Sutskever, *etal*, NIPS 2014]

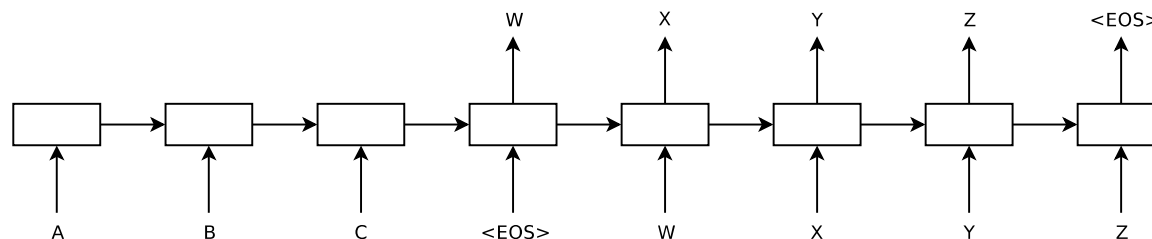
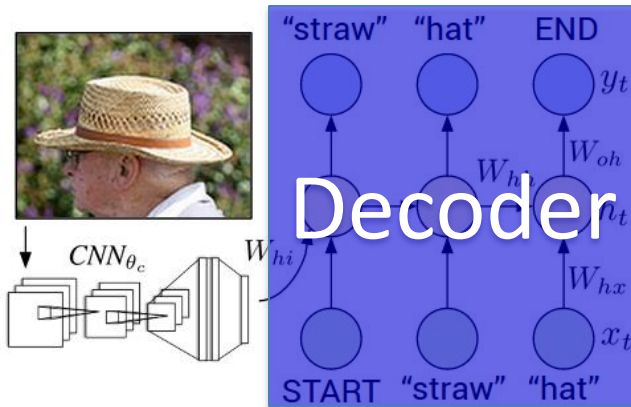


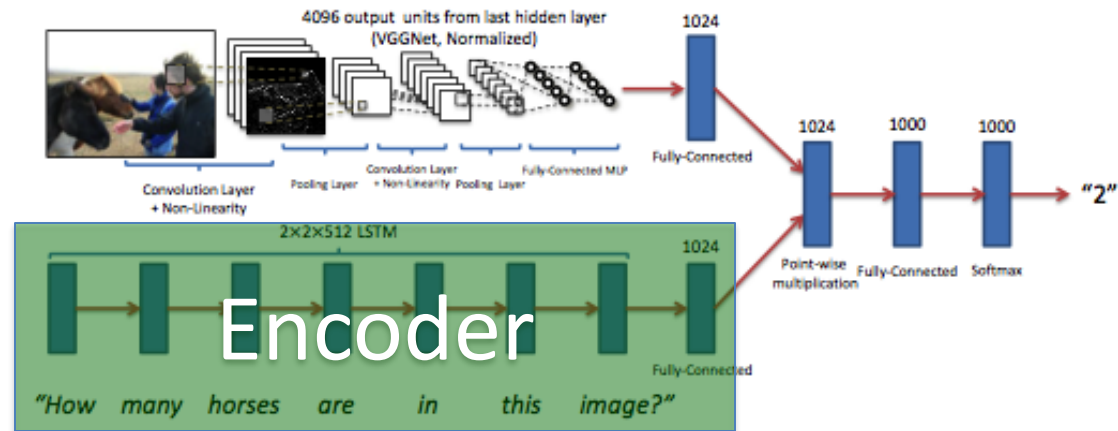
Image captioning

[Karpathy, *etal*, CVPR 2015]



Visual Question Answering

[Antol, *etal*, ICCV 2015]



Machine Translation

[Ilya Sutskever, *etal*, NIPS 2014]

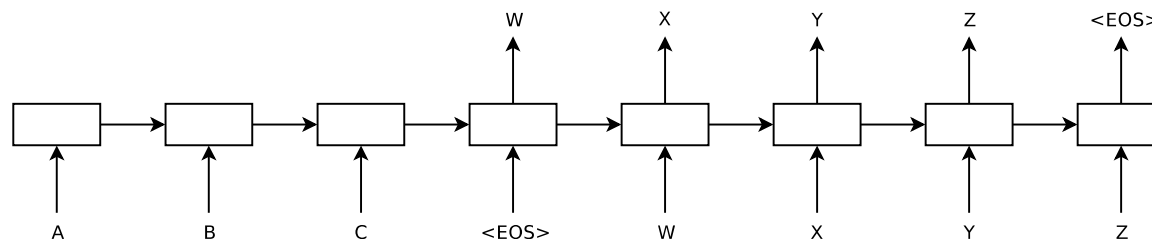
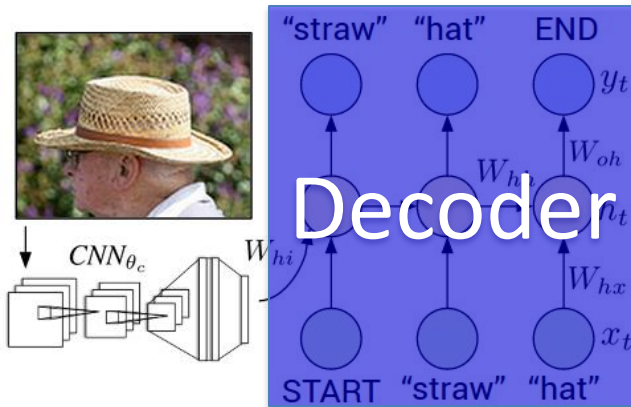


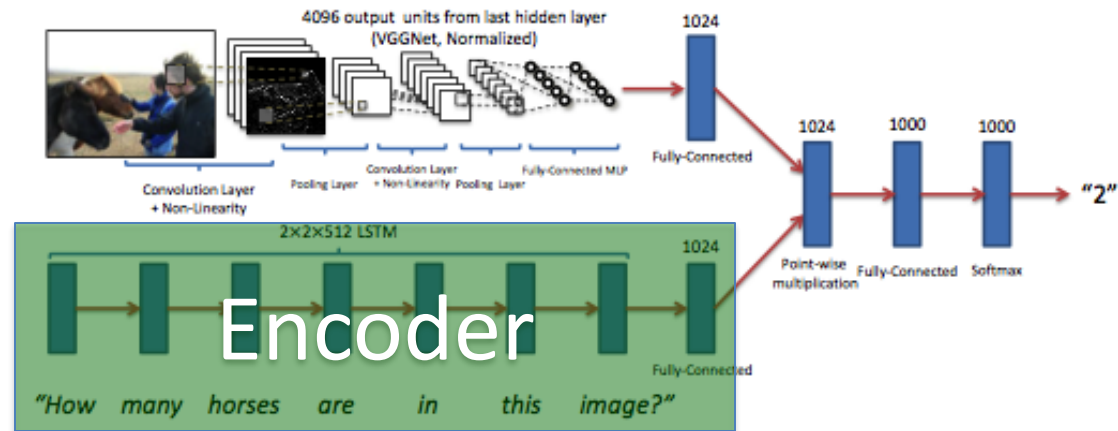
Image captioning

[Karpathy, *etal*, CVPR 2015]



Visual Question Answering

[Antol, *etal*, ICCV 2015]



Machine Translation

[Ilya Sutskever, *etal*, NIPS 2014]

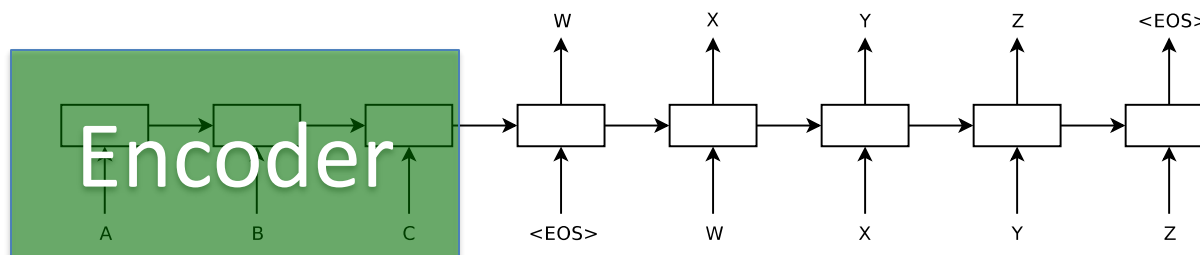
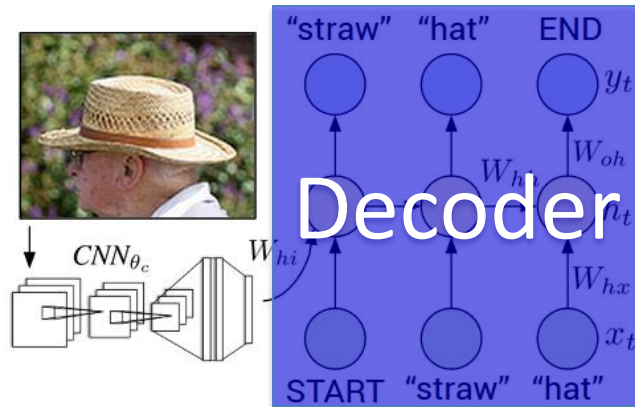


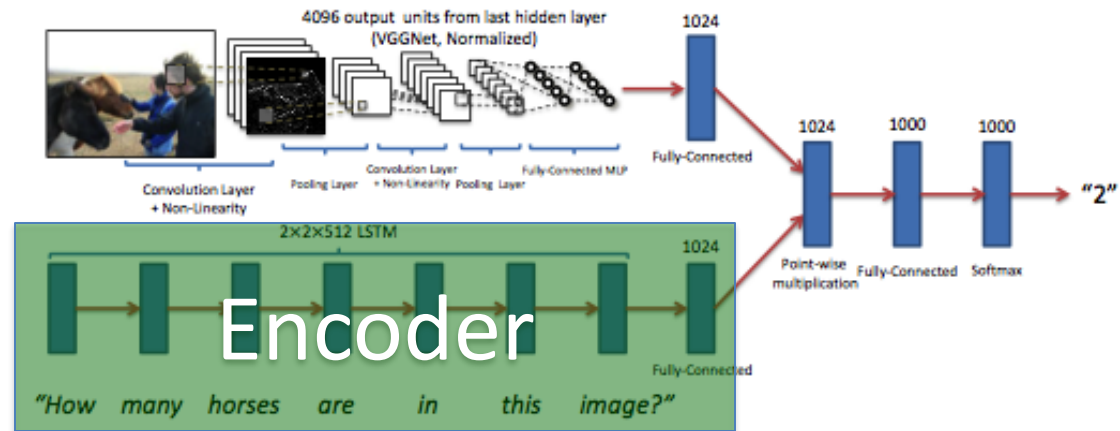
Image captioning

[Karpathy, *etal*, CVPR 2015]



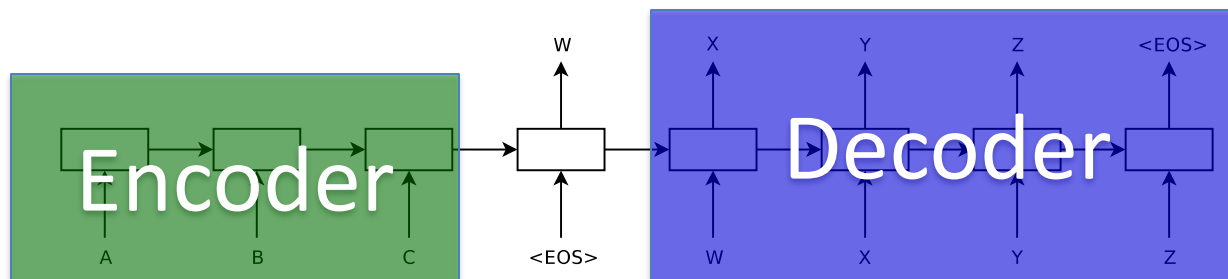
Visual Question Answering

[Antol, *etal*, ICCV 2015]



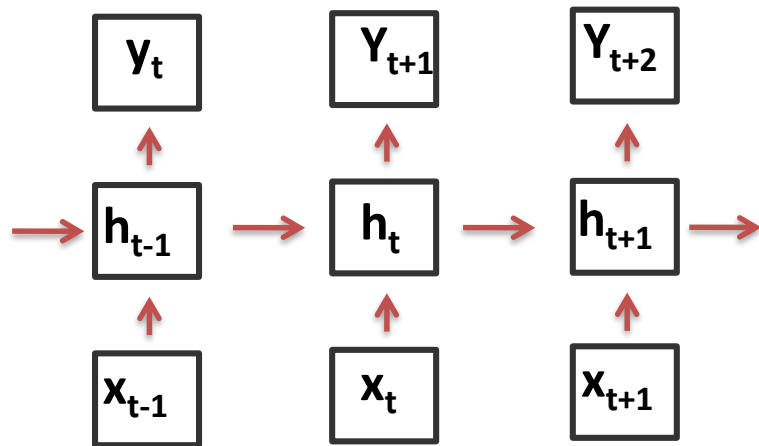
Machine Translation

[Ilya Sutskever, *etal*, NIPS 2014]



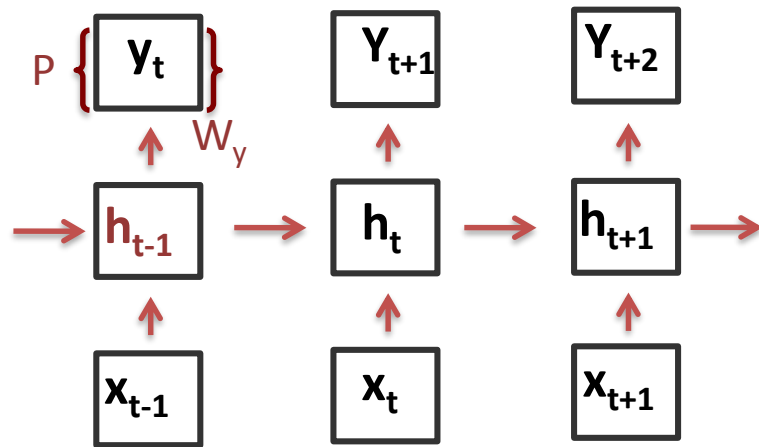
Uni-RNNs vs. Bi-RNNs

Uni-RNNs vs. Bi-RNNs



(a) Unidirectional RNNs

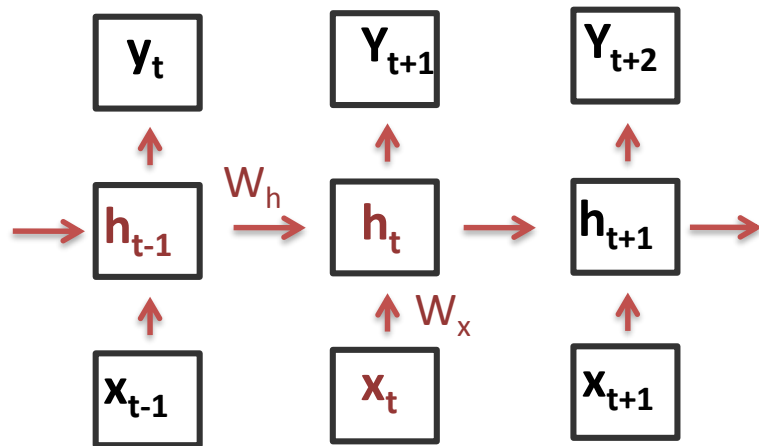
Uni-RNNs vs. Bi-RNNs



(a) Unidirectional RNNs

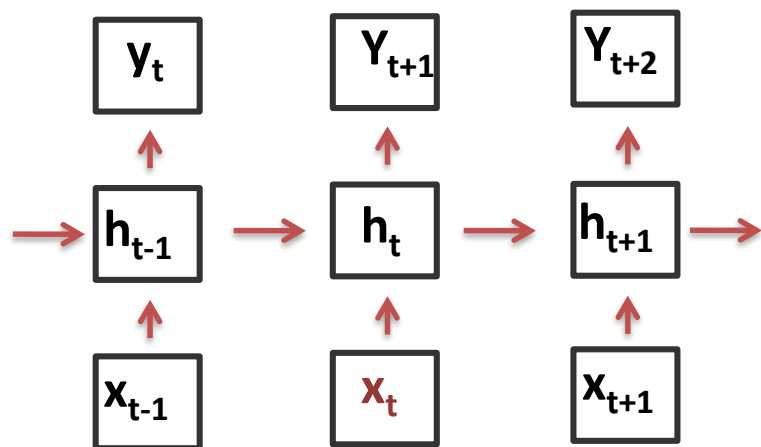
$$p(y_t | X_{[1:t-1]}) = \phi(W_y h_{t-1} + b_y)$$

Uni-RNNs vs. Bi-RNNs

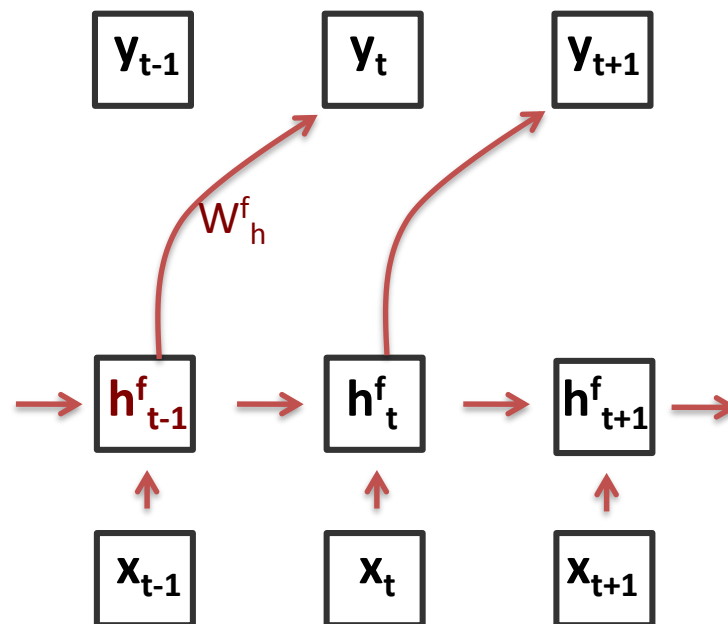


(a) Unidirectional RNNs

$$h_t = \tanh(W_x x_t + W_h h_{t-1} + b_h)$$

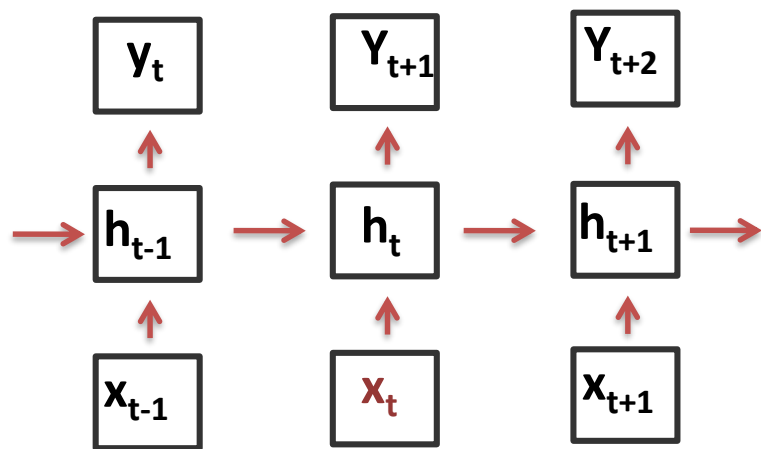


(a) Unidirectional RNNs

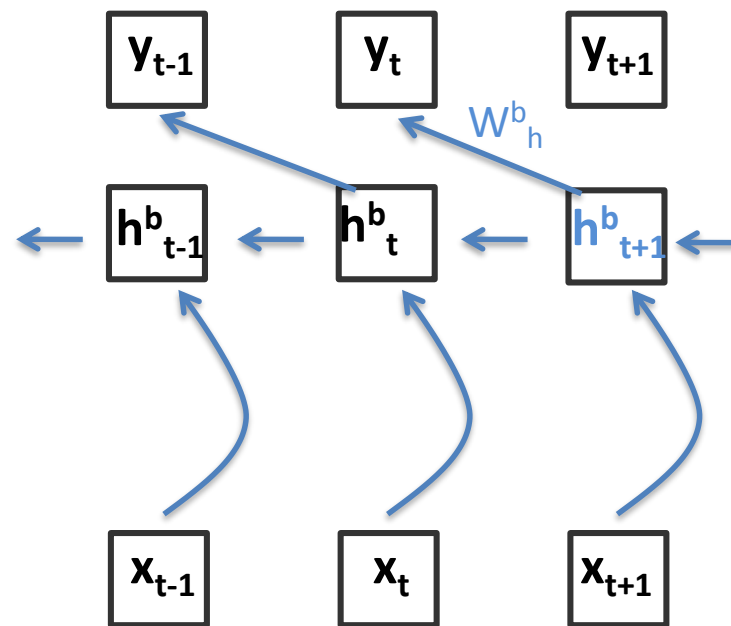


(b) Bidirectional RNNs

$$p(y_t | X_{[1:T] \setminus t}) = \phi(\textcolor{red}{W}_y^f h_{t-1}^f + \textcolor{blue}{W}_y^b h_{t+1}^b + b_y)$$

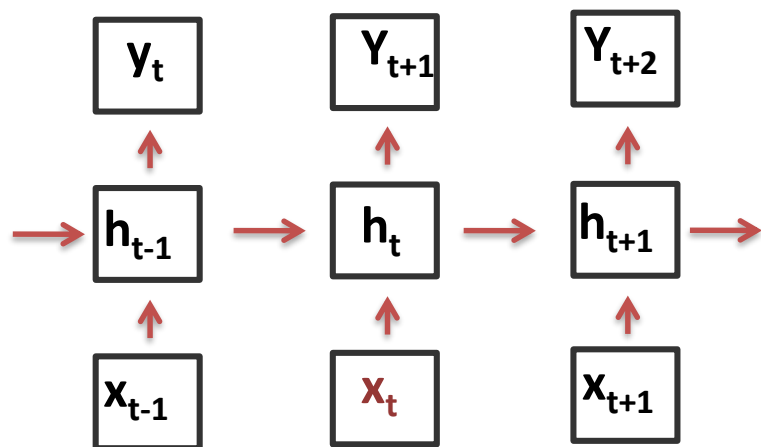


(a) Unidirectional RNNs

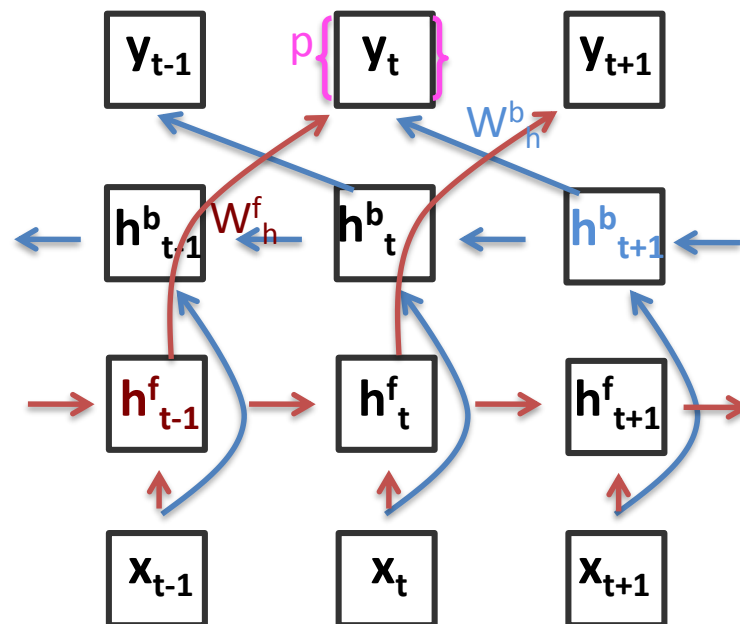


(b) Bidirectional RNNs

$$p(y_t | X_{[1:T] \setminus t}) = \phi(\textcolor{red}{W}_y^f h_{t-1}^f + \textcolor{blue}{W}_y^b h_{t+1}^b + b_y)$$



(a) Unidirectional RNNs



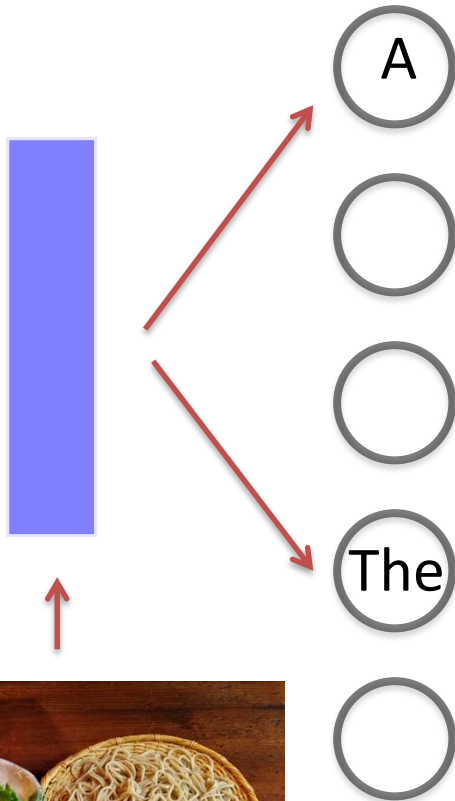
(b) Bidirectional RNNs

$$p(y_t | X_{[1:T] \setminus t}) = \phi(\textcolor{red}{W}_y^f h_{t-1}^f + \textcolor{blue}{W}_y^b h_{t+1}^b + b_y)$$

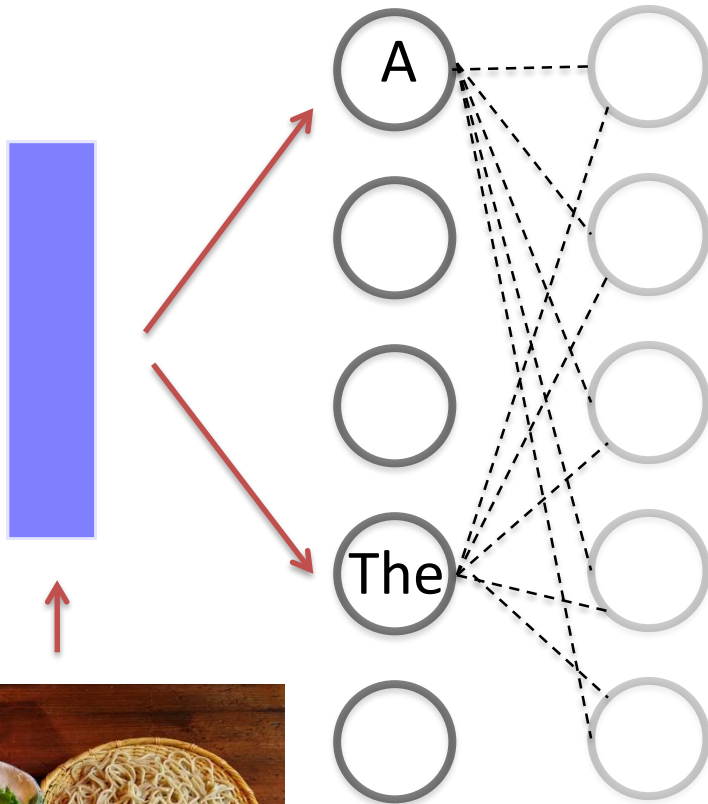
Left-to-right Beam Search



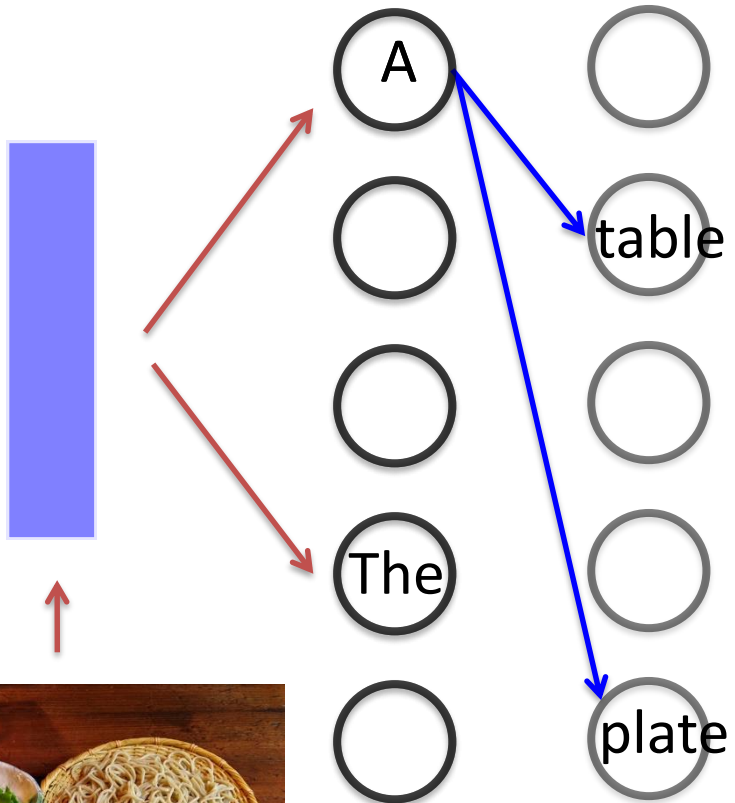
Left-to-right Beam Search



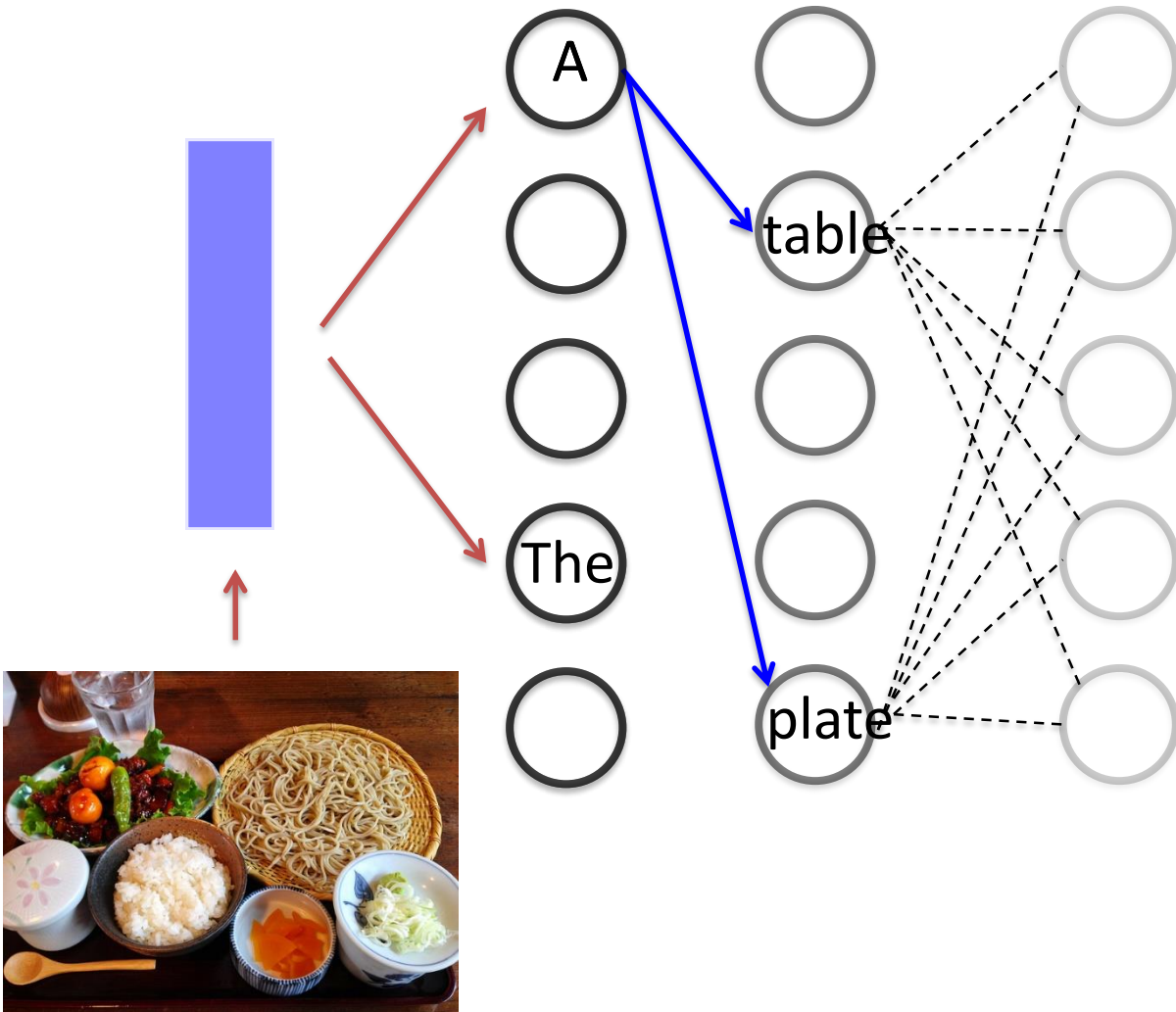
Left-to-right Beam Search



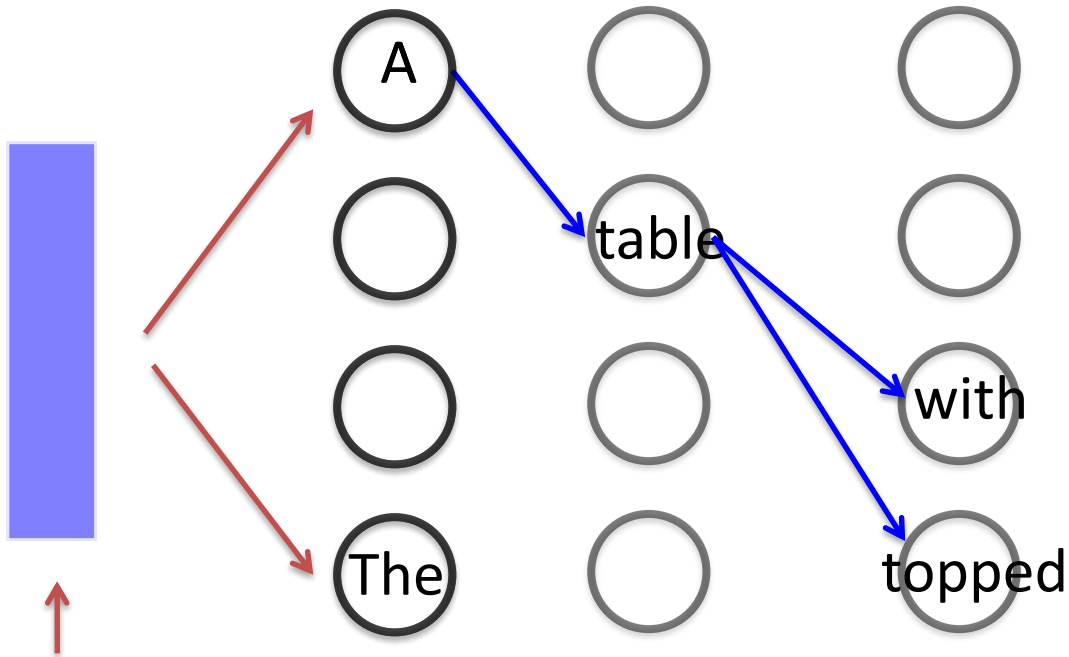
Left-to-right Beam Search



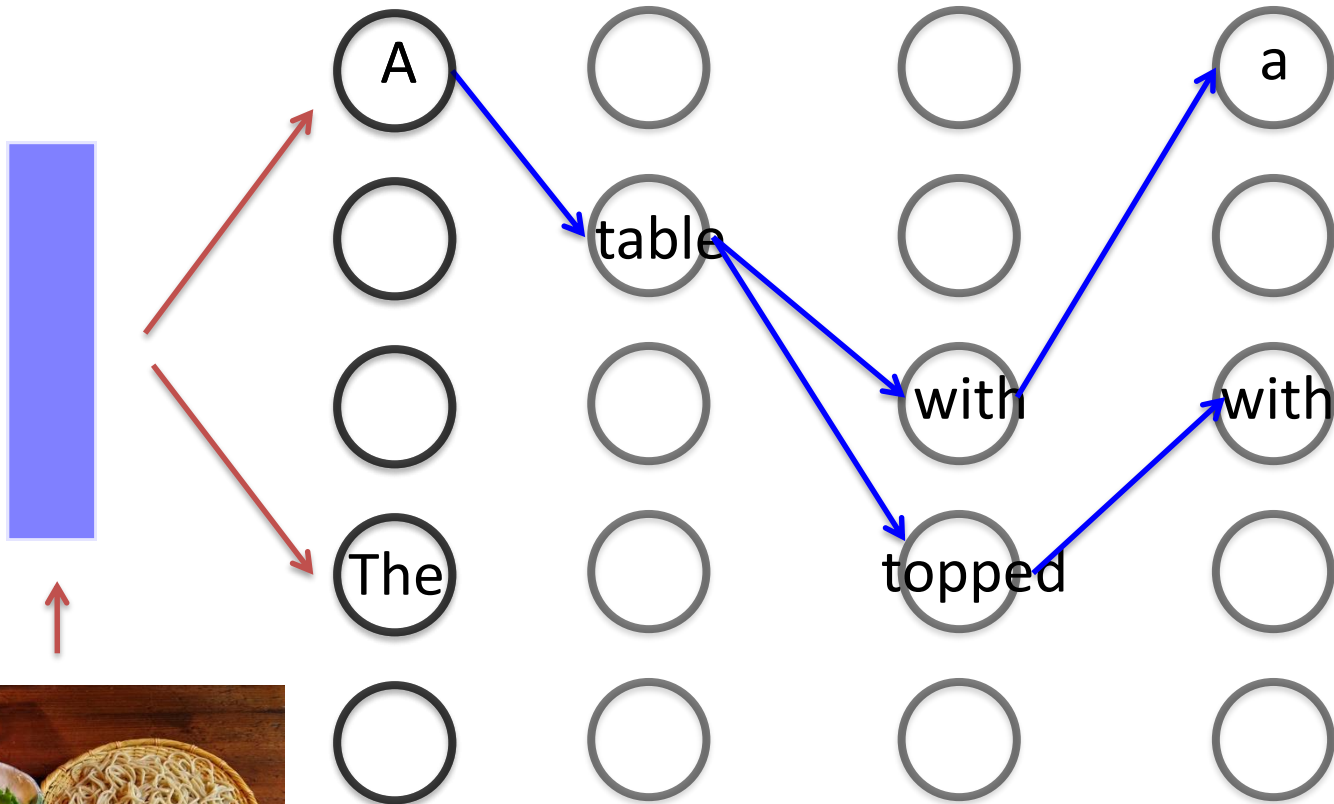
Left-to-right Beam Search



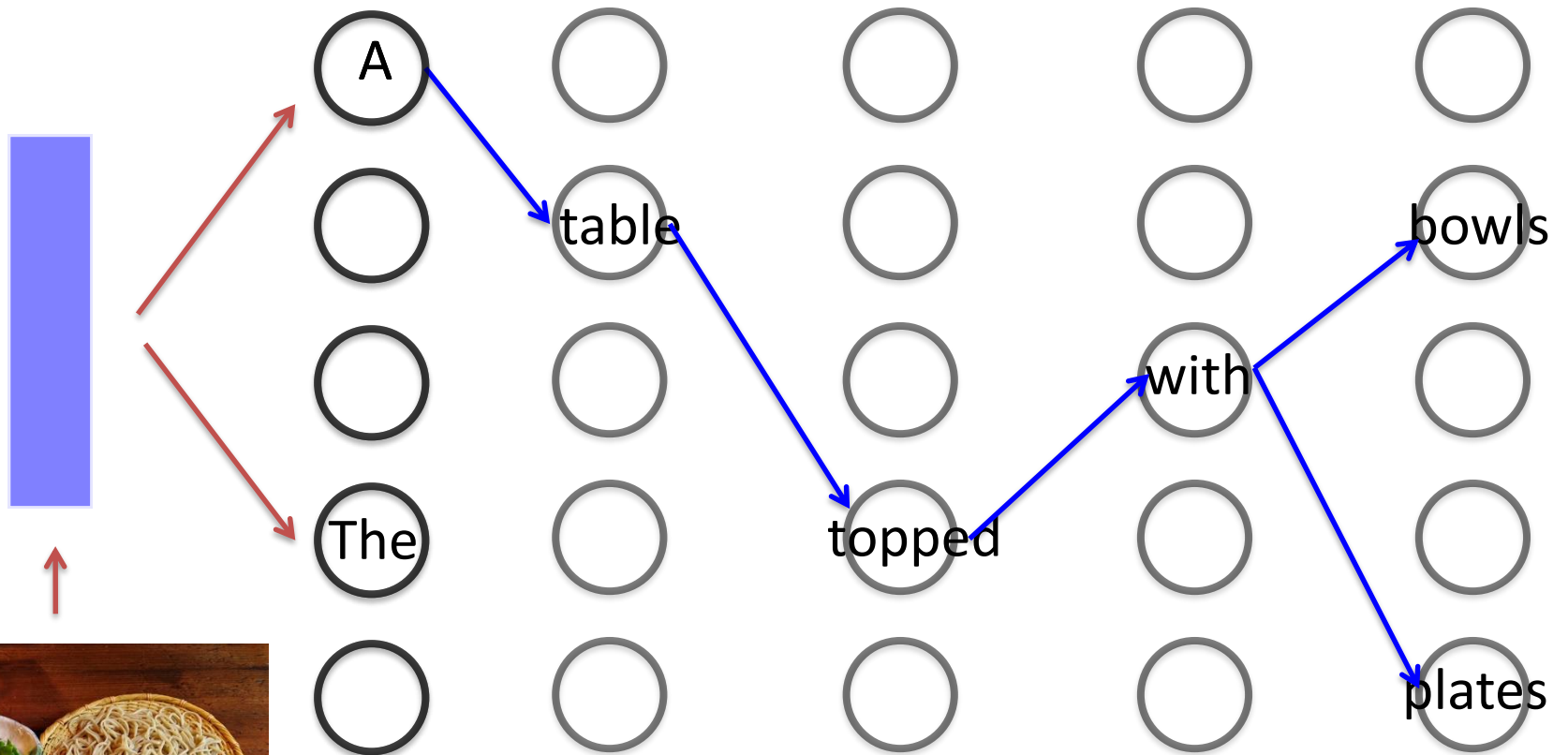
Left-to-right Beam Search



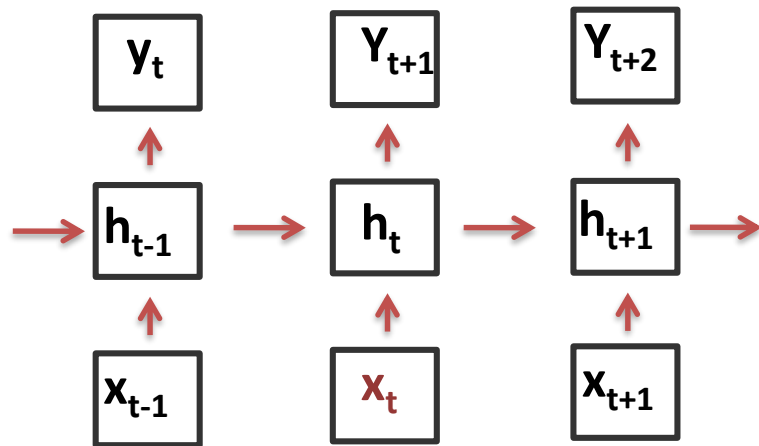
Left-to-right Beam Search



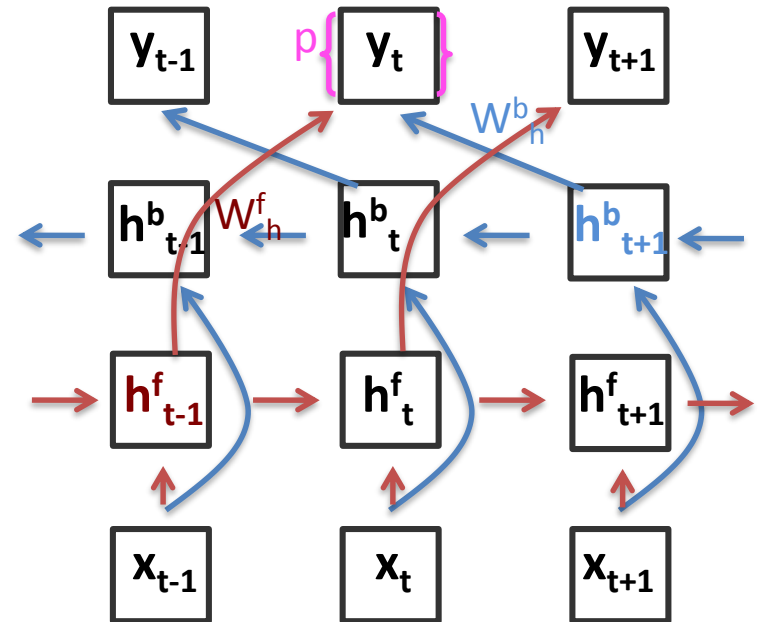
Left-to-right Beam Search



Left-to-right Beam Search



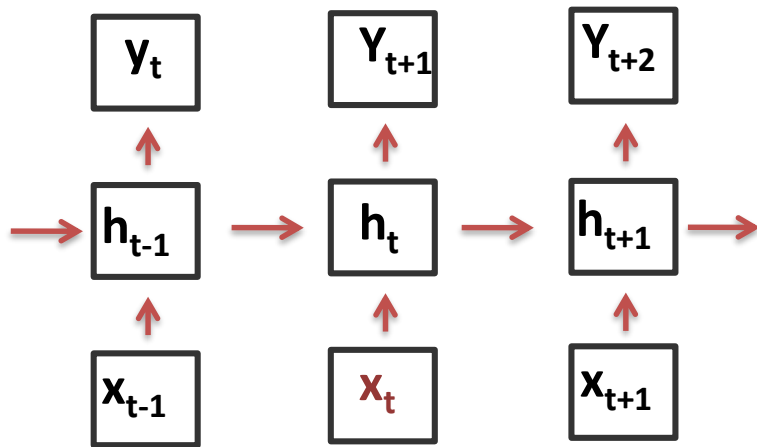
(a) Unidirectional RNNs



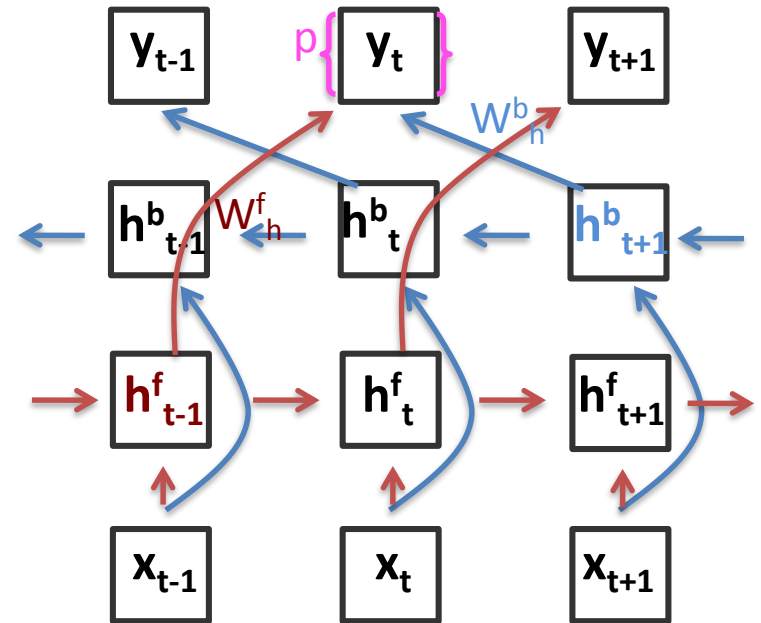
(b) Bidirectional RNNs

$$p(y_t | X_{[1:T] \setminus t}) = \phi(\textcolor{red}{W}_y^f h_{t-1}^f + \textcolor{blue}{W}_y^b h_{t+1}^b + b_y)$$

Left-to-right Beam Search



(a) Unidirectional RNNs



(b) Bidirectional RNNs


$$p(y_t | X_{[1:T] \setminus t}) = \phi(\textcolor{red}{W}_y^f h_{t-1}^f + \textcolor{blue}{W}_y^b h_{t+1}^b + b_y)$$


 Future variables

Inference in Bi-RNNs

Inference in Bi-RNNs


Fill-in-the-blank Image Captioning

 **COCO**
Common Objects in Context


[cocodataset@outlook.com](#)

[Home](#) [People](#) [Explore](#) [Dataset](#) [External](#)

[Overview](#) [Challenges](#) [Download](#) [Evaluate](#) [Leaderboard](#)




The man at bat readies to swing at the pitch while the umpire looks on.



A large bus sitting next to a very tall building.

Inference in Bi-RNNs


Fill-in-the-blank Image Captioning

 **COCO**
Common Objects in Context


[cocodataset@outlook.com](#)

[Home](#) [People](#) [Explore](#) [Dataset](#) [External](#)

[Overview](#) [Challenges](#) [Download](#) [Evaluate](#) [Leaderboard](#)



The man at bat readies to swing at the pitch while the umpire looks on.



A large bus sitting next to a very tall building.

Visual Madlibs

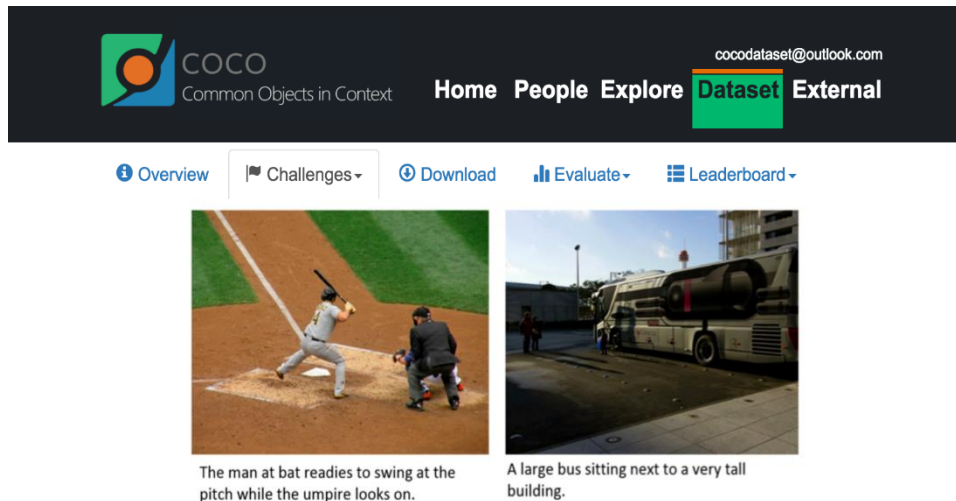


The people are eating cake at the dining table.
The people are serving the cake.


Visual Madlibs:
Fill-In-The-Blank And Question-Answering


Inference in Bi-RNNs

Fill-in-the-blank Image Captioning



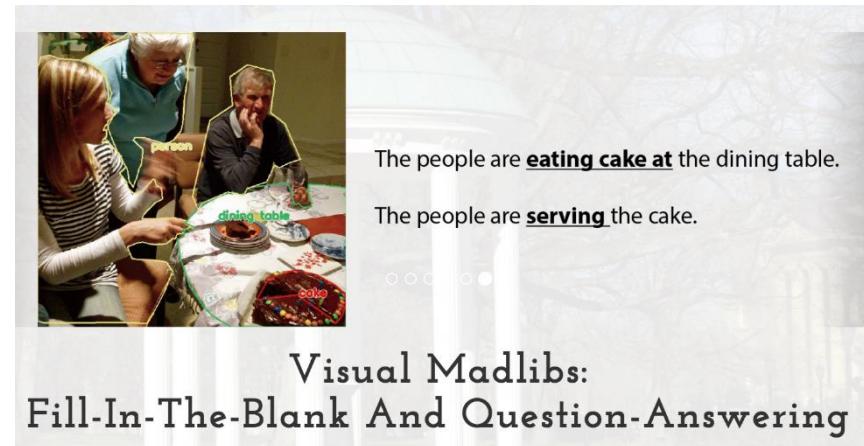
The screenshot shows the COCO (Common Objects in Context) dataset website. The header includes the COCO logo, the text "Common Objects in Context", the email "cocodataset@outlook.com", and navigation links: "Home", "People", "Explore", "Dataset", and "External". Below the header are links for "Overview", "Challenges", "Download", "Evaluate", and "Leaderboard". Two image examples are shown with captions:

- 

The man at bat readies to swing at the pitch while the umpire looks on.
- 

A large bus sitting next to a very tall building.

Visual Madlibs



The Visual Madlibs interface shows a photo of three people at a dining table. The captions are:

- The people are eating cake at the dining table.
- The people are serving the cake.

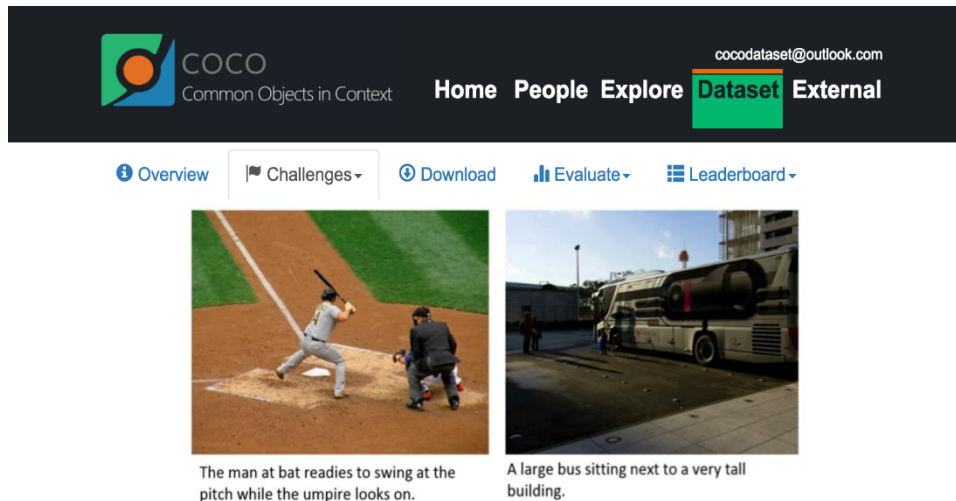
Visual Madlibs:
Fill-In-The-Blank And Question-Answering

Image Completion/Impainting



Inference in Bi-RNNs

Fill-in-the-blank Image Captioning

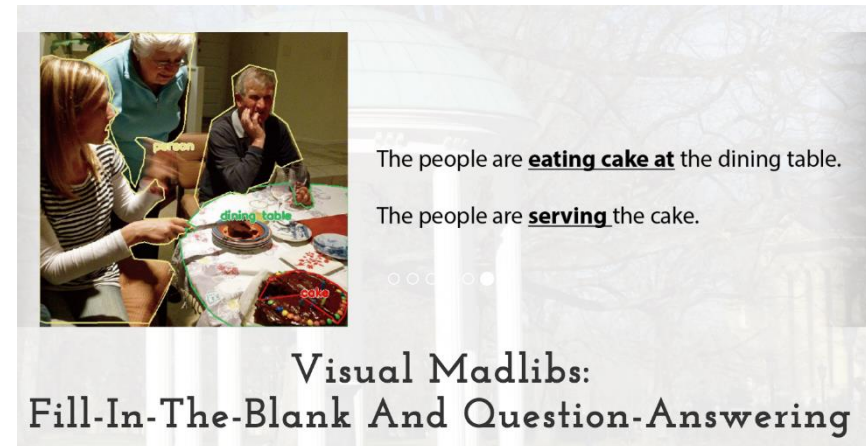


The screenshot shows the COCO dataset website interface. At the top, there's a navigation bar with links: Home, People, Explore, Dataset (highlighted in green), and External. Below this is a secondary navigation bar with links: Overview, Challenges, Download, Evaluate, and Leaderboard. The main content area displays two image captioning examples. The first example shows a baseball player at bat with the caption: "The man at bat readies to swing at the pitch while the umpire looks on." The second example shows a large bus next to a tall building with the caption: "A large bus sitting next to a very tall building."

Image Completion/Impainting

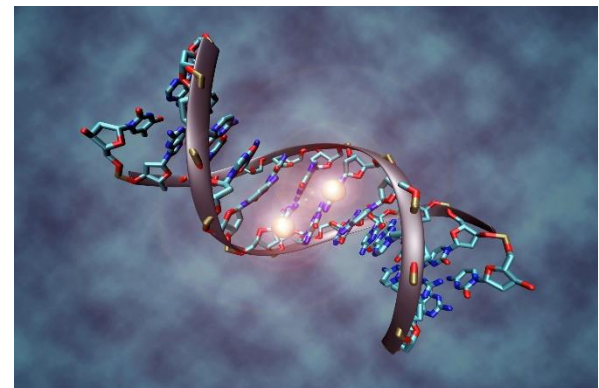


Visual Madlibs



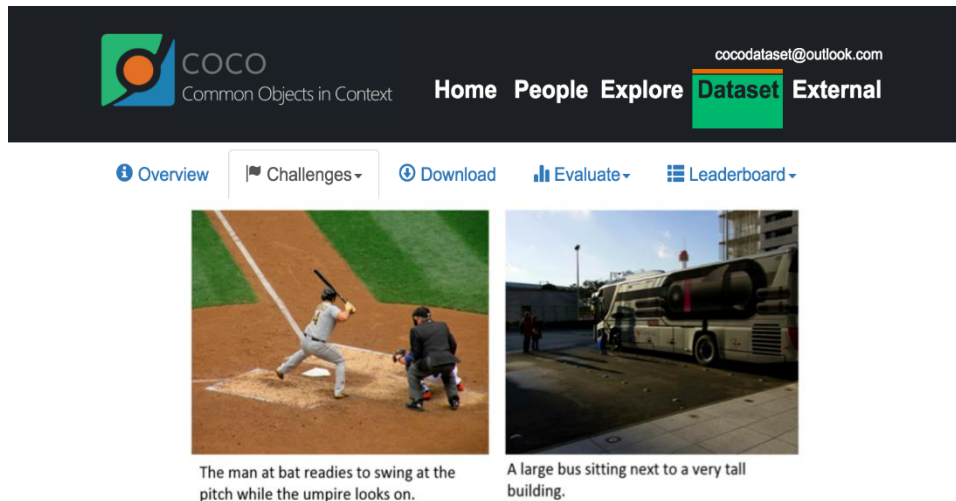
The image shows a Visual Madlibs interface. It features a photograph of three people sitting at a dining table. Overlaid on the photo are labels for "person" and "dining table". To the right of the photo, there are two sentences with blanks for user input: "The people are eating cake at the dining table." and "The people are serving the cake." Below the sentences are five small circles, with the third one filled, indicating the current position in the sequence. At the bottom, the text reads: "Visual Madlibs: Fill-In-The-Blank And Question-Answering".

Genome Sequencing





Inference in Bi-RNNs

Fill-in-the-blank Image Captioning




The screenshot shows the COCO dataset website. The header includes the COCO logo, the text "Common Objects in Context", the email "cocodataset@outlook.com", and navigation links: "Home", "People", "Explore", "Dataset", and "External". Below the header are links for "Overview", "Challenges", "Download", "Evaluate", and "Leaderboard". Two image captioning examples are shown:

- 

The man at bat readies to swing at the pitch while the umpire looks on.
- 

A large bus sitting next to a very tall building.

Visual Madlibs



The Visual Madlibs interface shows a photo of people at a table. The captions are:

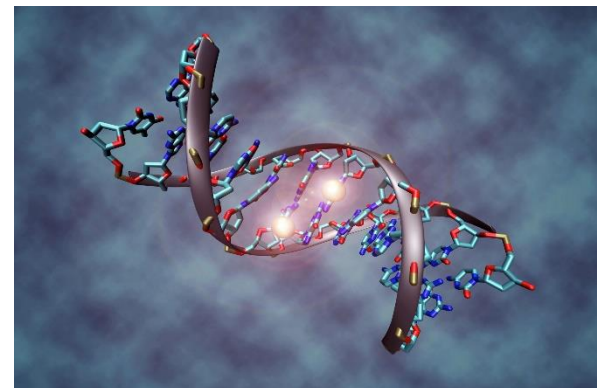
- The people are eating cake at the dining table.
- The people are serving the cake.

Visual Madlibs:
Fill-In-The-Blank And Question-Answering

Image Completion/Impainting



Genome Sequencing





A girl and a dog are balancing on a wind board



A girl and

a wind board



URNN-f: A girl and a dog are in the a wind board



URNN-b: A girl and sitting in the water with a wind board



BiRNN+BSCD: A girl and a dog are sitting on a wind board



A girl in a room full of books wearing a long red tie



A girl in

_a long red tie



URNN-f: A girl in a dress shirt and tie standing a long red tie



URNN-b: A girl in a woman is wearing a long red tie



BiRNN+BSCD: A girl in a white dress shirt is holding a long red tie

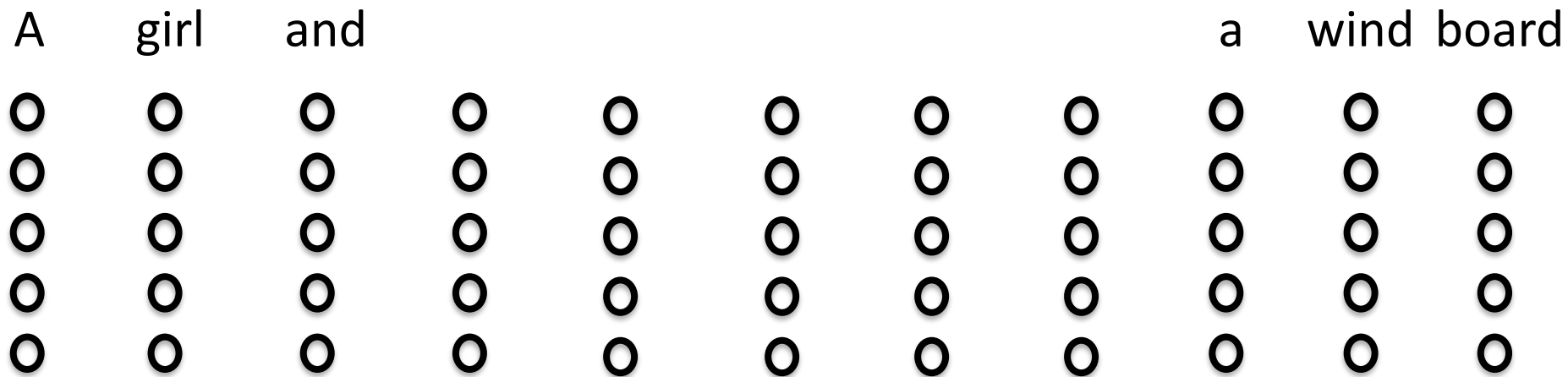
Inference in Bi-RNNs

Contributions:

Beam-based Top-B MAP Inference algorithm for Bi-RNNs

Left-to-right BS in Uni-RNN-f

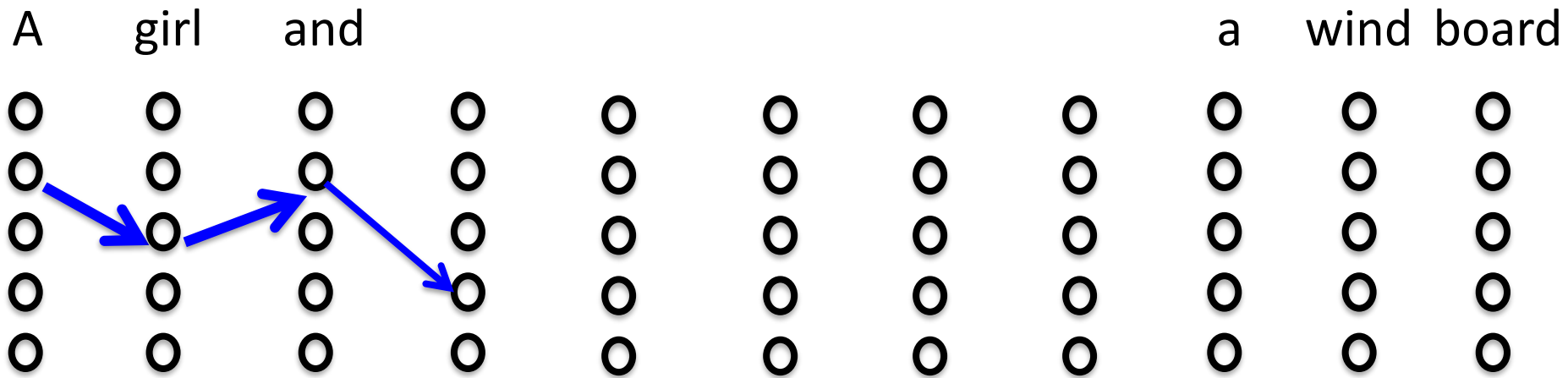
A girl and a wind board



The diagram illustrates a left-to-right beam search process. It consists of the words 'A', 'girl', 'and', 'a', 'wind', and 'board' arranged horizontally. Below each word, there are five vertically stacked circles, representing a beam of size 5. This visualizes the search space at each time step, where only the top 5 scoring partial sequences are expanded to the next step.

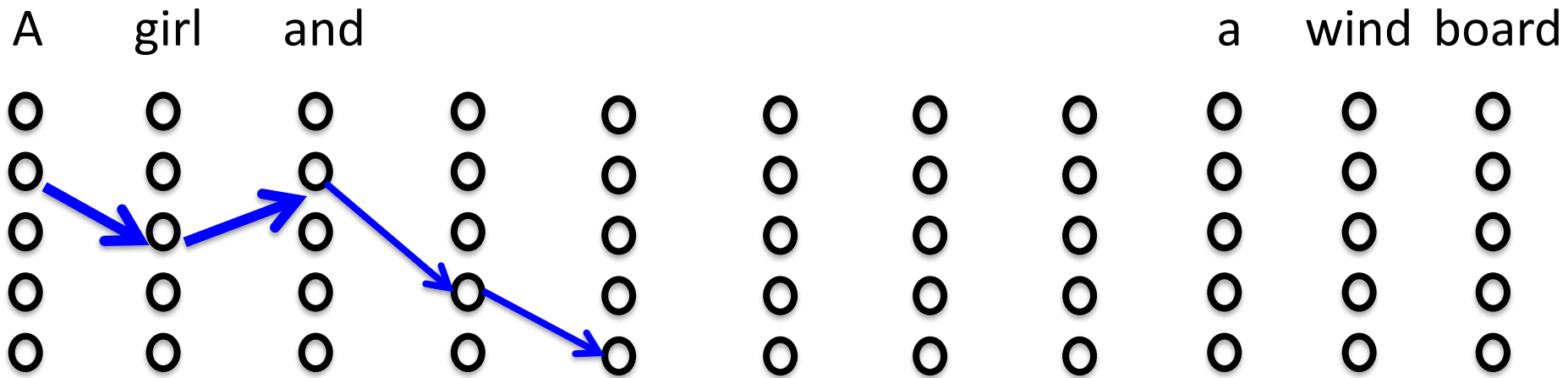
$$S^f(t) = S^f(t-1) + \log p(y_t | y_{[1:t-1]})$$

Left-to-right BS in Uni-RNN-f



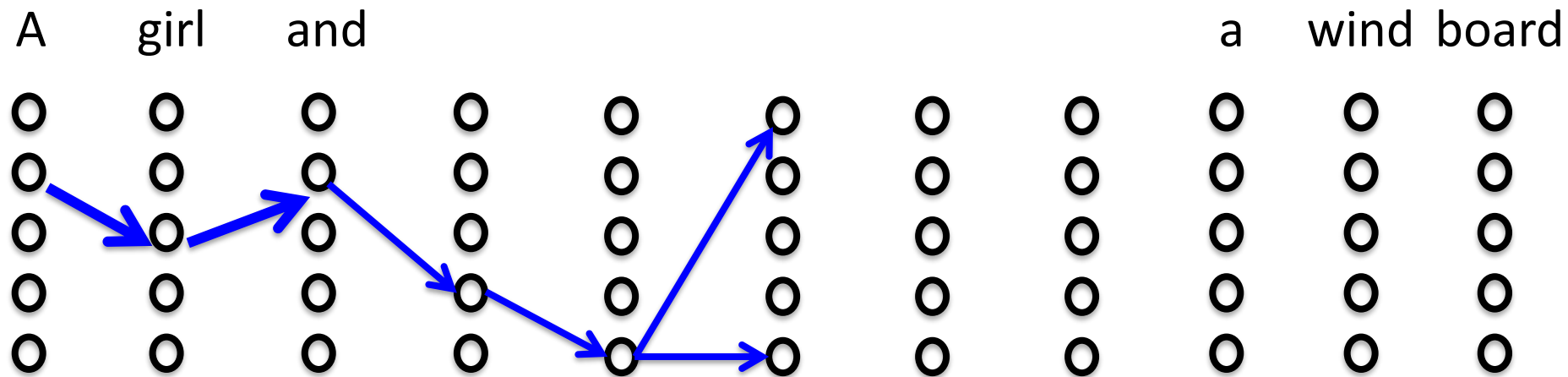
$$S^f(t) = S^f(t-1) + \log p(y_t | y_{[1:t-1]})$$

Left-to-right BS in Uni-RNN-f



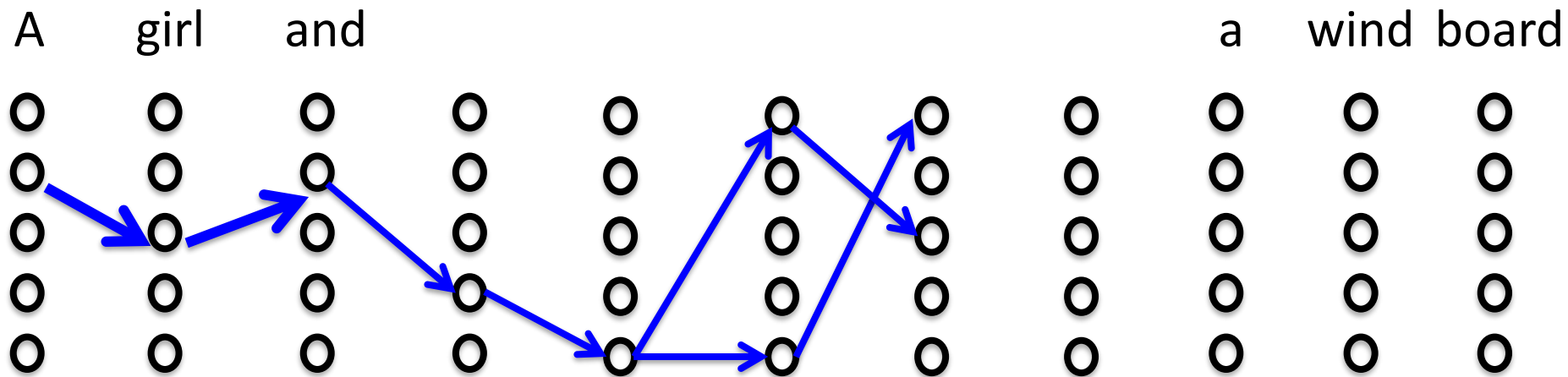
$$S^f(t) = S^f(t-1) + \log p(y_t | y_{[1:t-1]})$$

Left-to-right BS in Uni-RNN-f



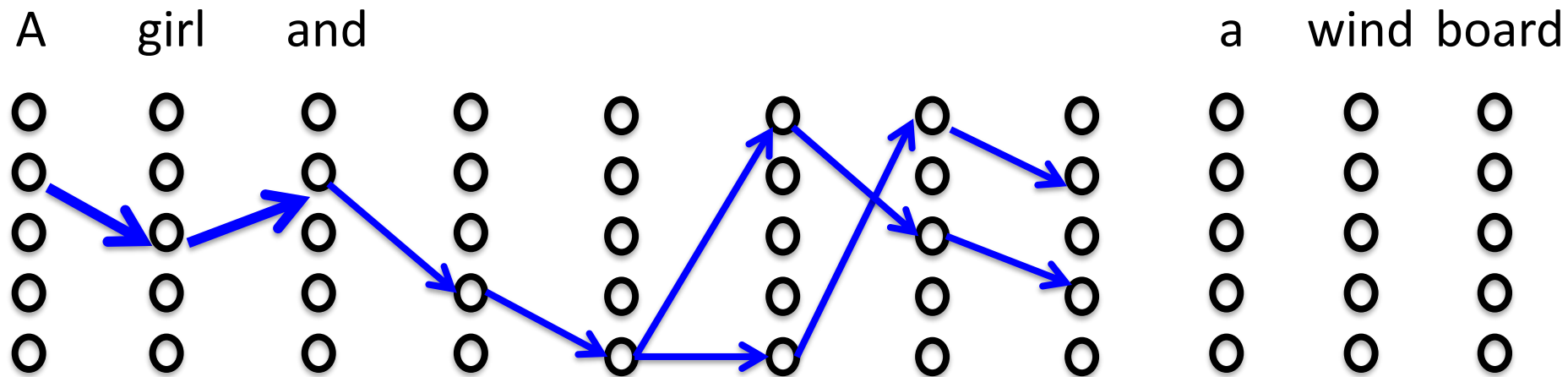
$$S^f(t) = S^f(t-1) + \log p(y_t | y_{[1:t-1]})$$

Left-to-right BS in Uni-RNN-f



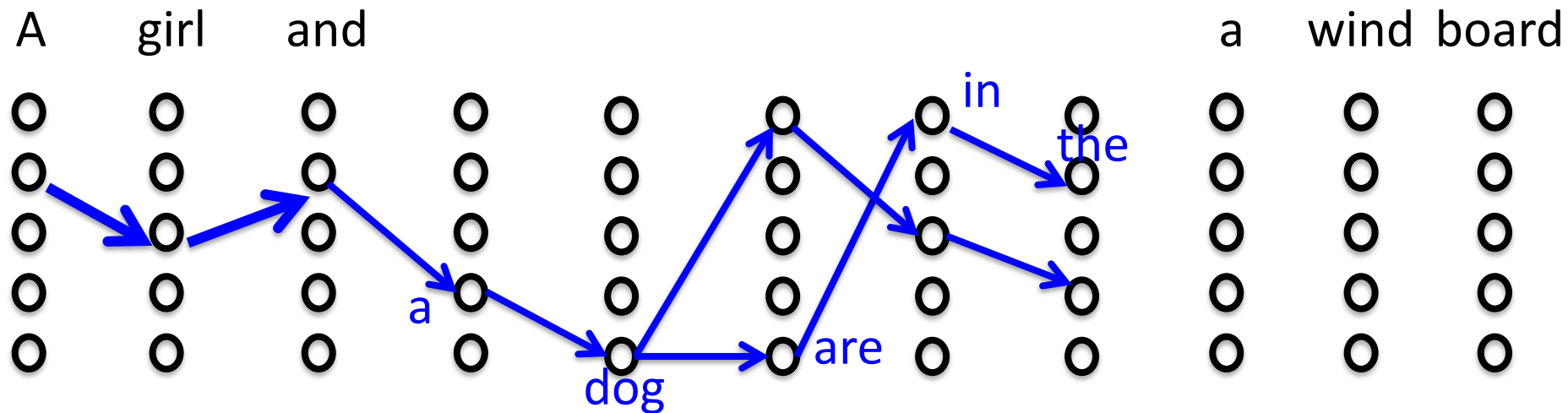
$$S^f(t) = S^f(t-1) + \log p(y_t | y_{[1:t-1]})$$

Left-to-right BS in Uni-RNN-f



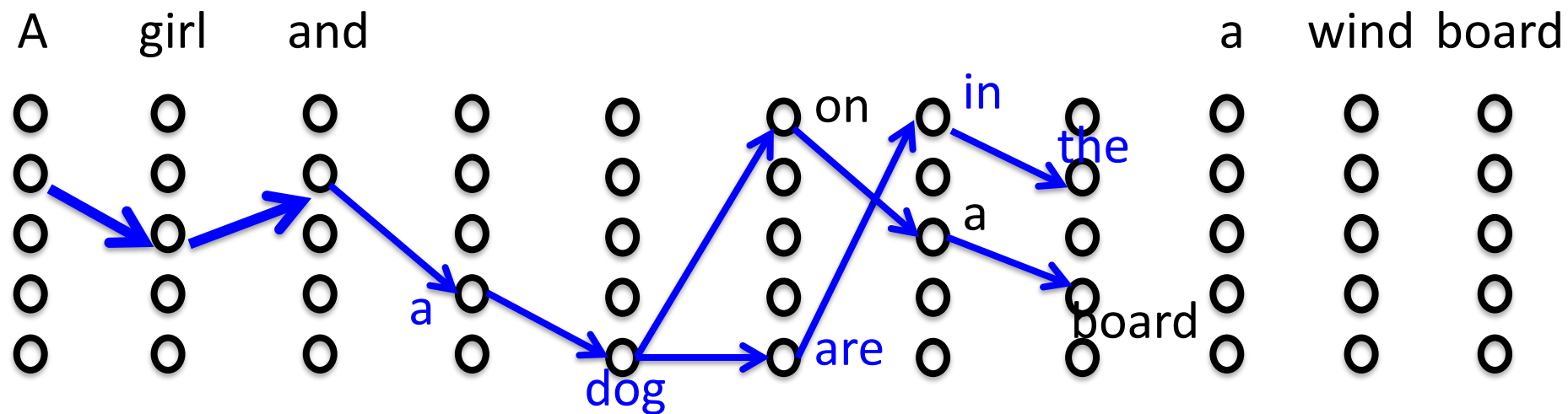
$$S^f(t) = S^f(t-1) + \log p(y_t | y_{[1:t-1]})$$

Left-to-right BS in Uni-RNN-f



$$S^f(t) = S^f(t-1) + \log p(y_t | y_{[1:t-1]})$$

Left-to-right BS in Uni-RNN-f



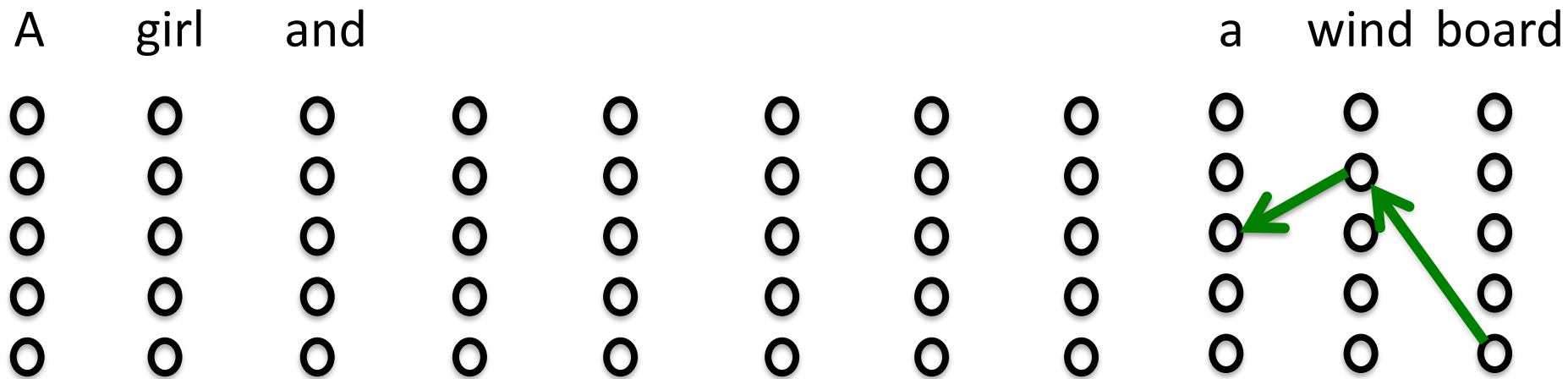
$$S^f(t) = S^f(t-1) + \log p(y_t | y_{[1:t-1]})$$

Right-to-left BS in Uni-RNN-b

A	girl	and						a	wind	board
○	○	○	○	○	○	○	○	○	○	○
○	○	○	○	○	○	○	○	○	○	○
○	○	○	○	○	○	○	○	○	○	○
○	○	○	○	○	○	○	○	○	○	○
○	○	○	○	○	○	○	○	○	○	○

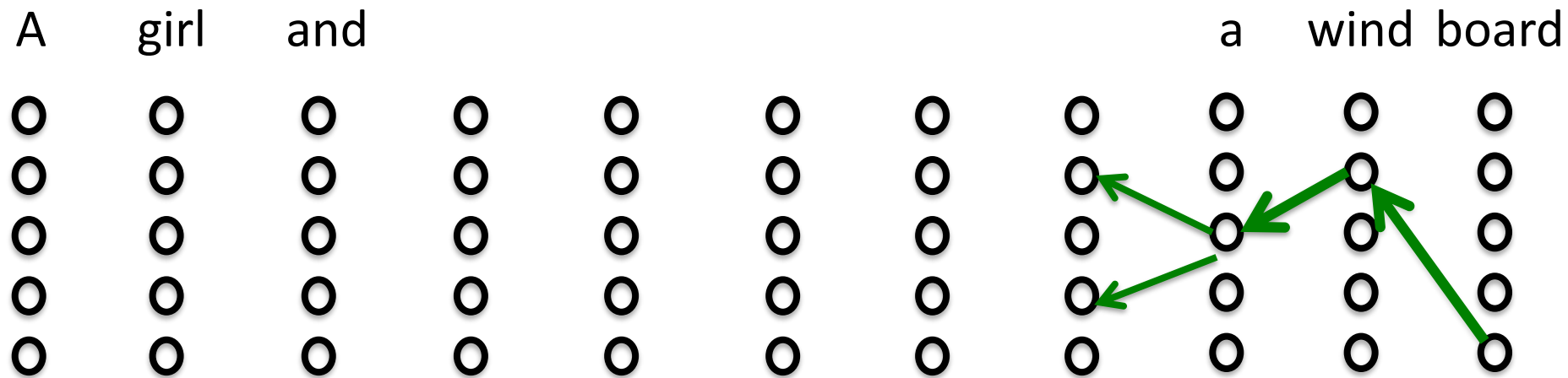
$$S^b(t) = S^b(t + 1) + \log p(y_t | y_{[t+1:T]})$$

Right-to-left BS in Uni-RNN-b



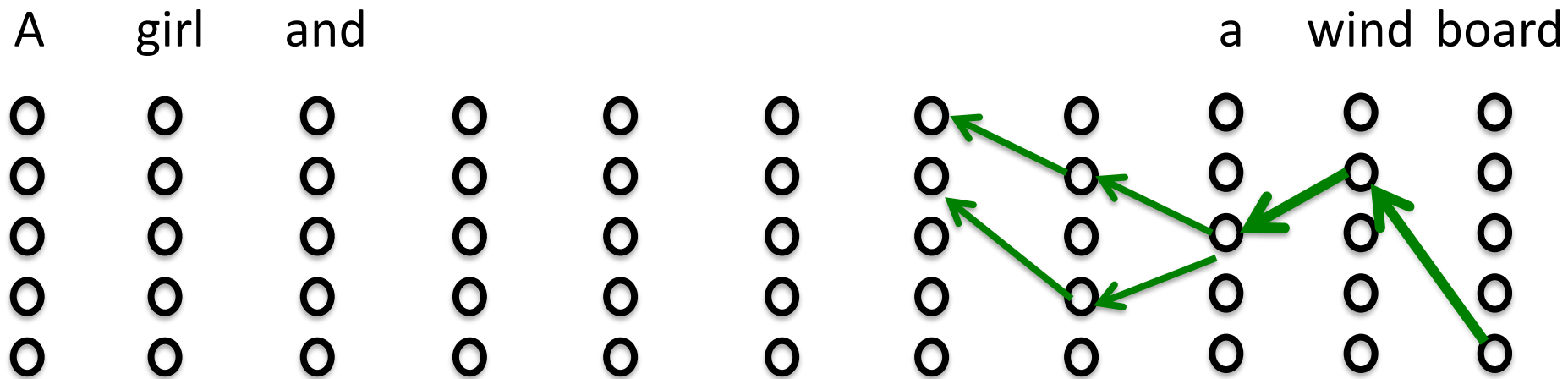
$$S^b(t) = S^b(t + 1) + \log p(y_t | y_{[t+1:T]})$$

Right-to-left BS in Uni-RNN-b



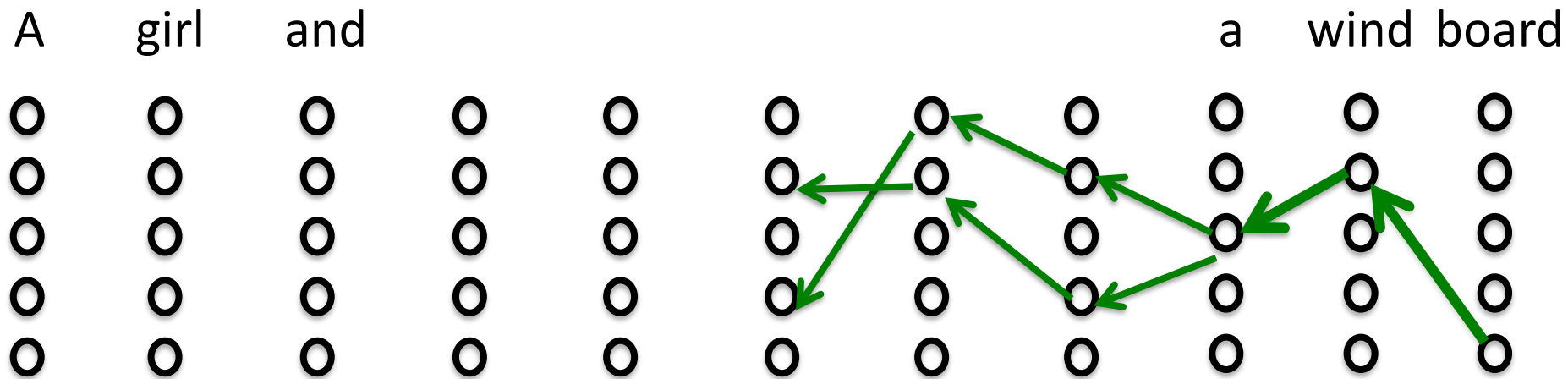
$$S^b(t)=S^b(t+1)+\log p(y_t|y_{[t+1:T]})$$

Right-to-left BS in Uni-RNN-b



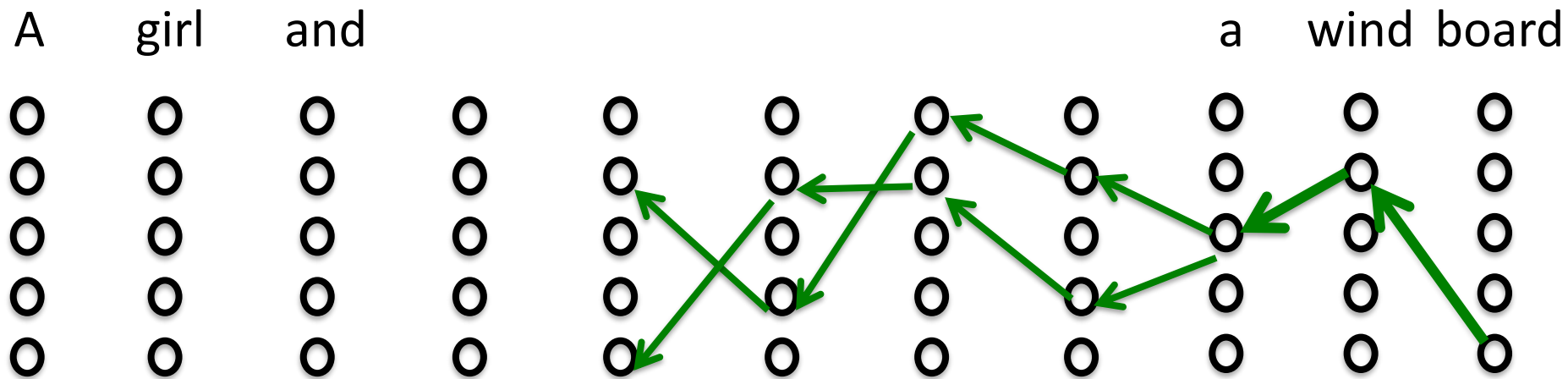
$$S^b(t)=S^b(t+1)+\log p(y_t|y_{[t+1:T]})$$

Right-to-left BS in Uni-RNN-b



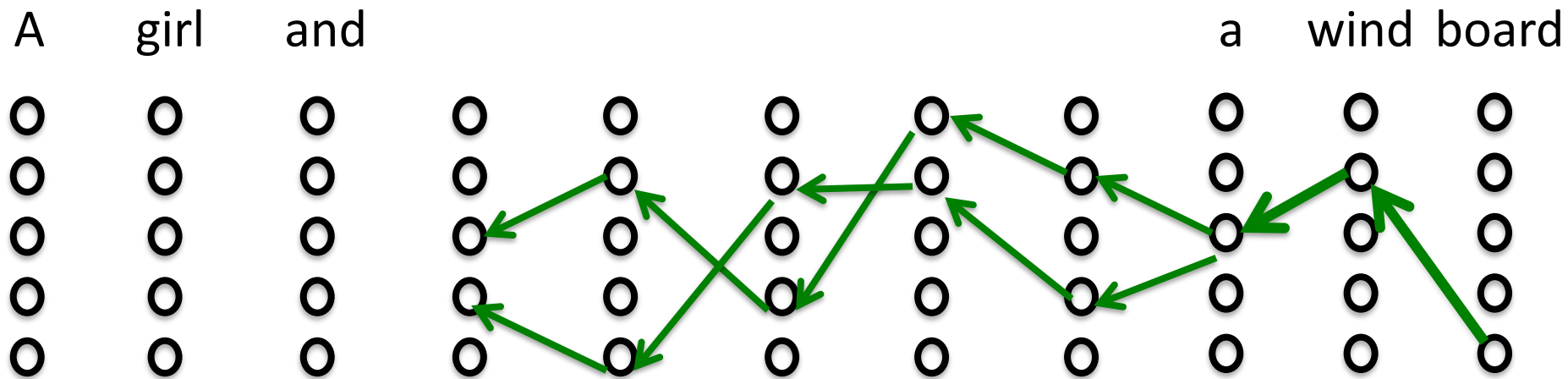
$$S^b(t)=S^b(t+1)+\log p(y_t|y_{[t+1:T]})$$

Right-to-left BS in Uni-RNN-b



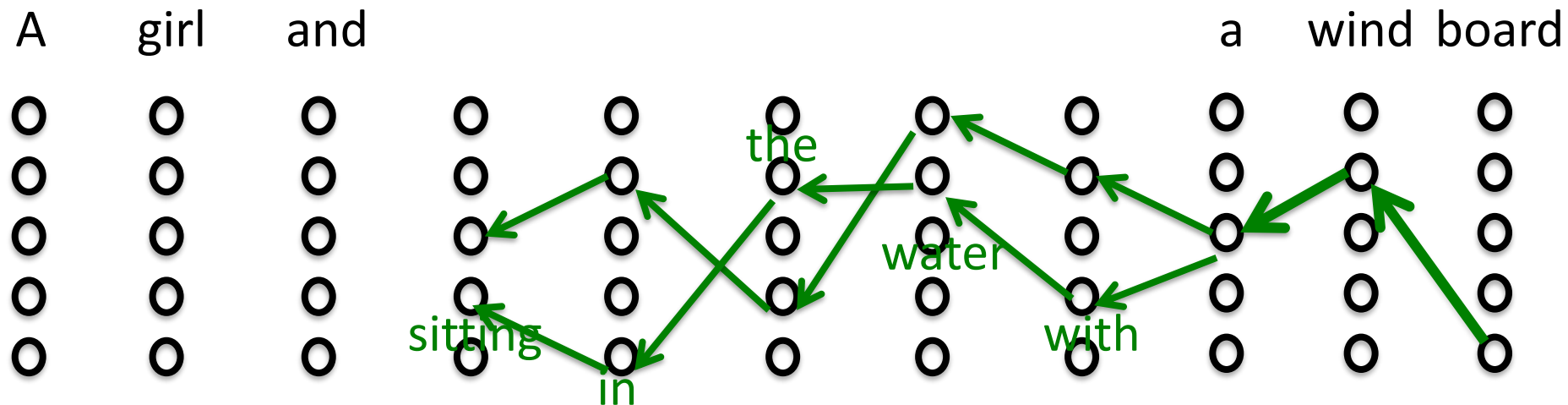
$$S^b(t)=S^b(t+1)+\log p(y_t|y_{[t+1:T]})$$

Right-to-left BS in Uni-RNN-b



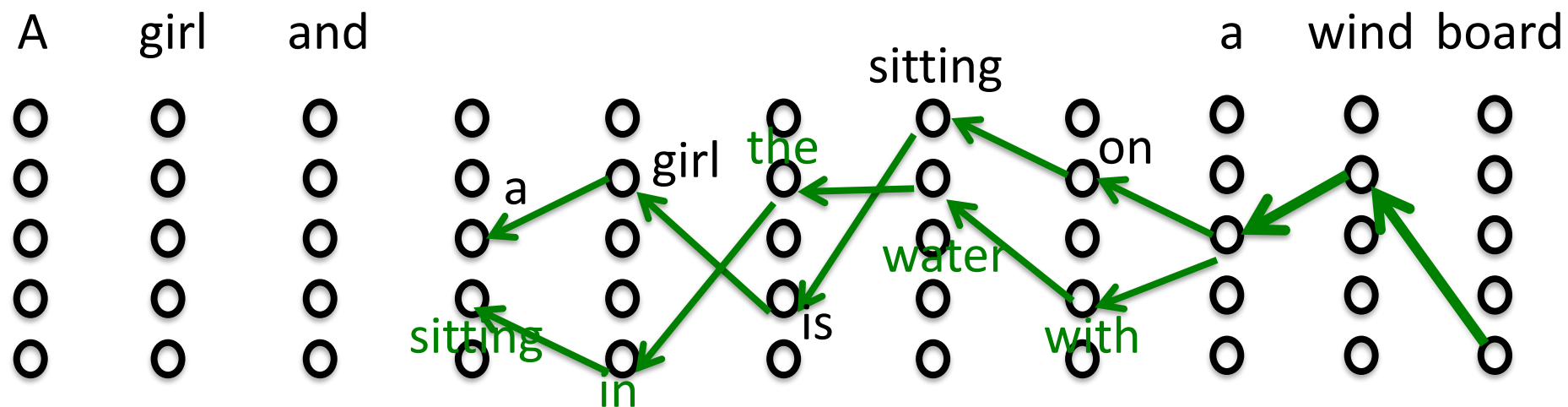
$$S^b(t)=S^b(t+1)+\log p(y_t|y_{[t+1:T]})$$

Right-to-left BS in Uni-RNN-b



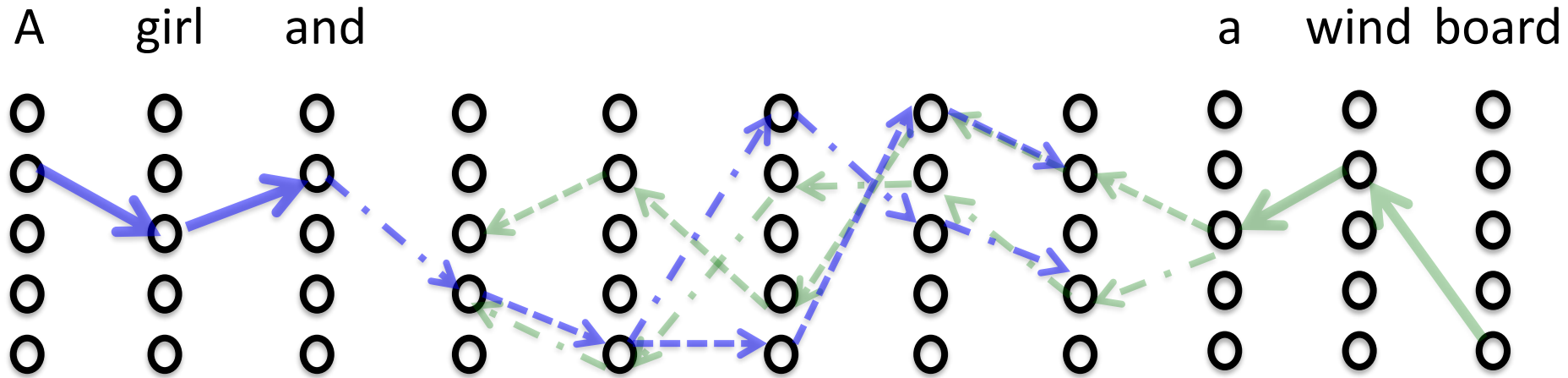
$$S^b(t) = S^b(t + 1) + \log p(y_t | y_{[t+1:T]})$$

Right-to-left BS in Uni-RNN-b



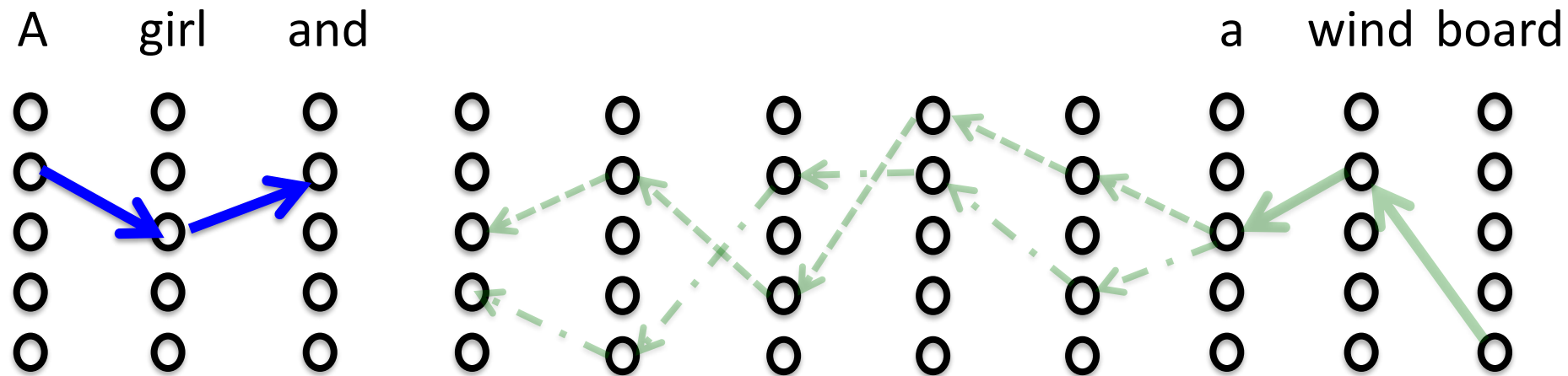
$$S^b(t) = S^b(t + 1) + \log p(y_t | y_{[t+1:T]})$$

Beam-Search Coordinate Descent (BSCD) in Bi-RNNs

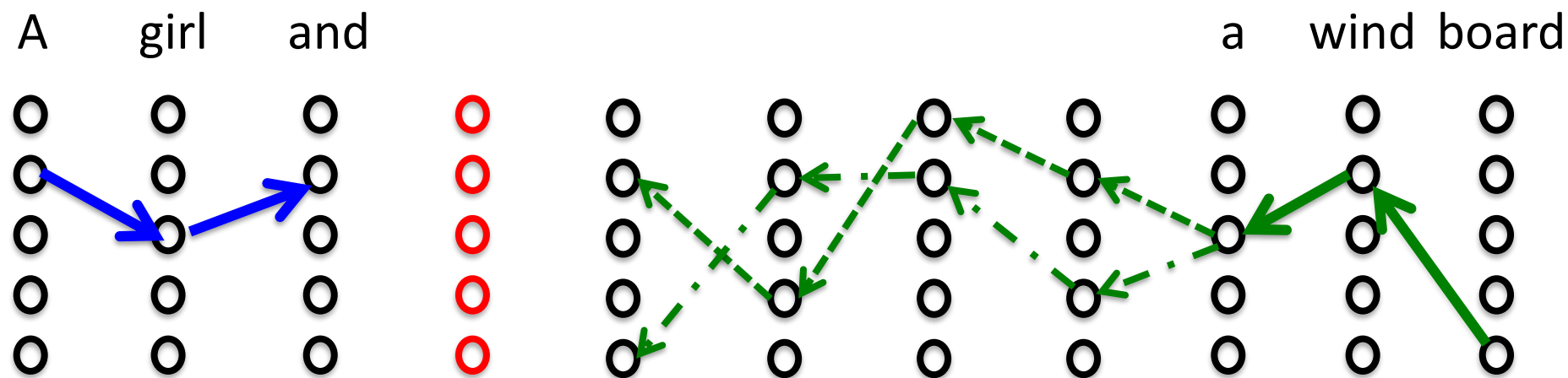


Initialize forward & backward beams using classical BS

Beam-Search Coordinate Descent (BSCD) in Bi-RNNs

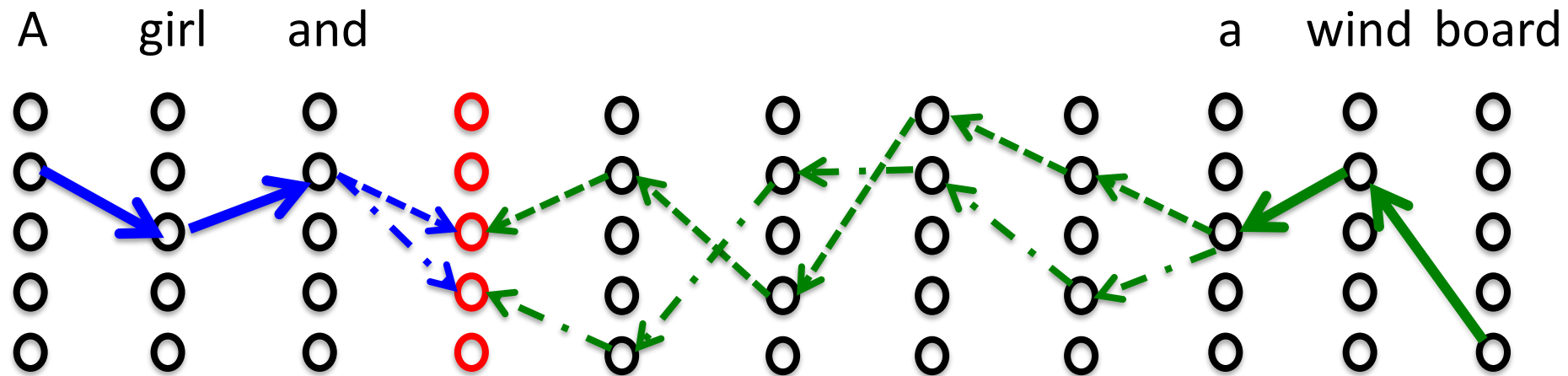


Beam-Search Coordinate Descent (BSCD) in Bi-RNNs

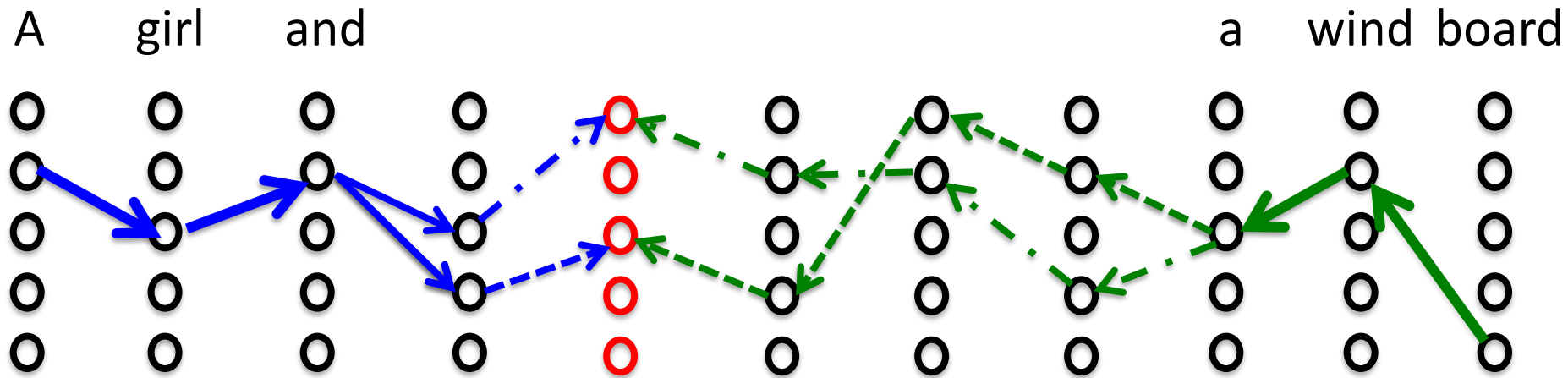


$$S(\leftarrow, \circ, \rightarrow) = S^f(t-1) + \log p(y_t | y_{t' \neq t}) + S^b(t+1)$$

Beam-Search Coordinate Descent (BSCD) in Bi-RNNs

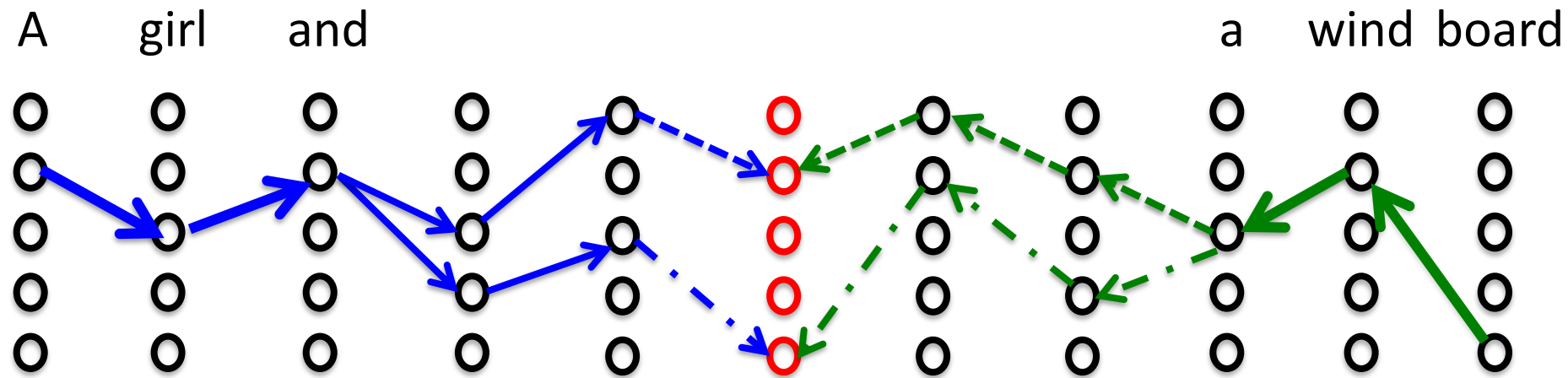


Beam-Search Coordinate Descent (BSCD) in Bi-RNNs

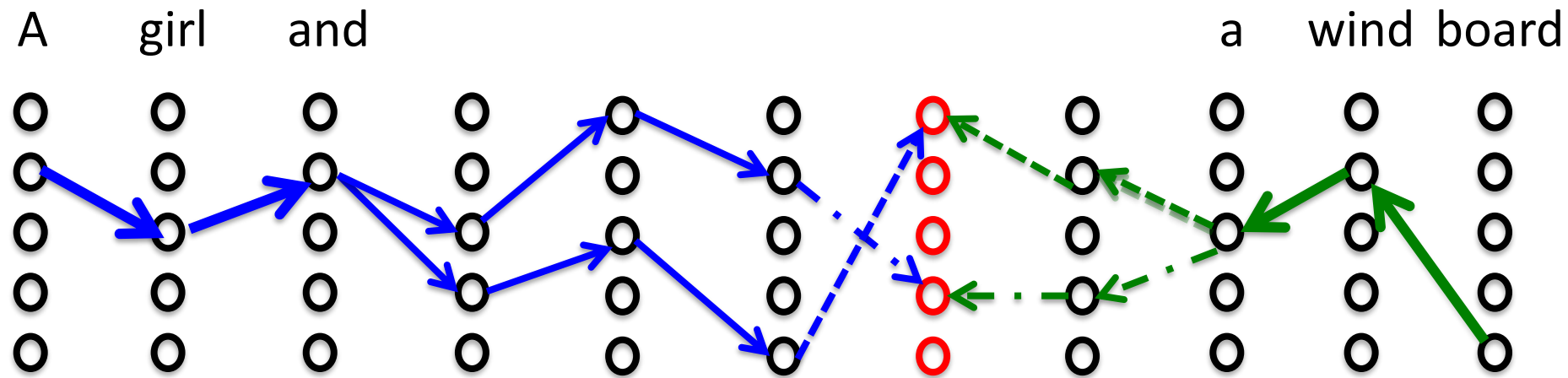


$$S^f(t) = S^f(t-1) + \log p(y_t | y_{t'} \neq t)$$

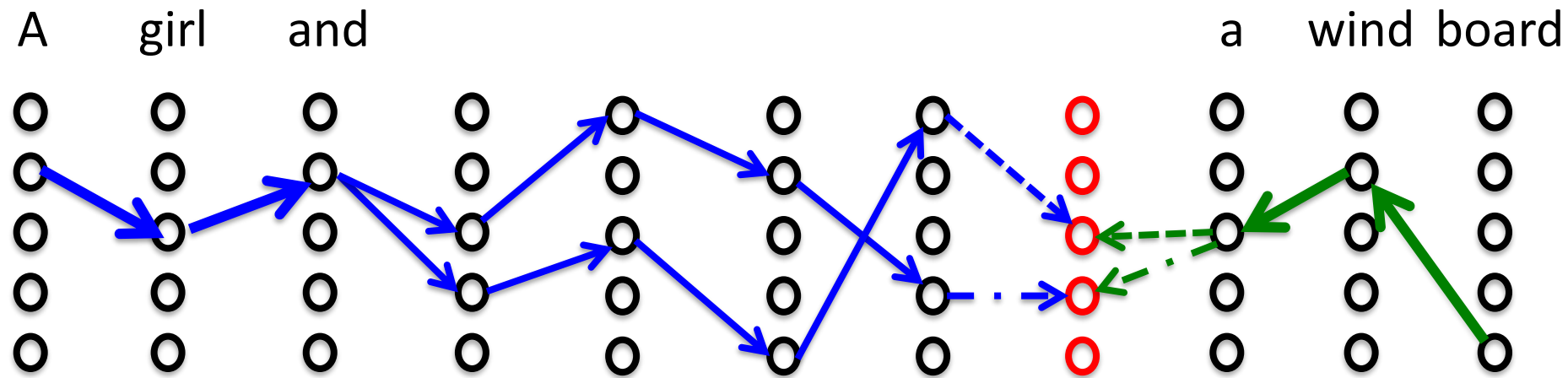
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



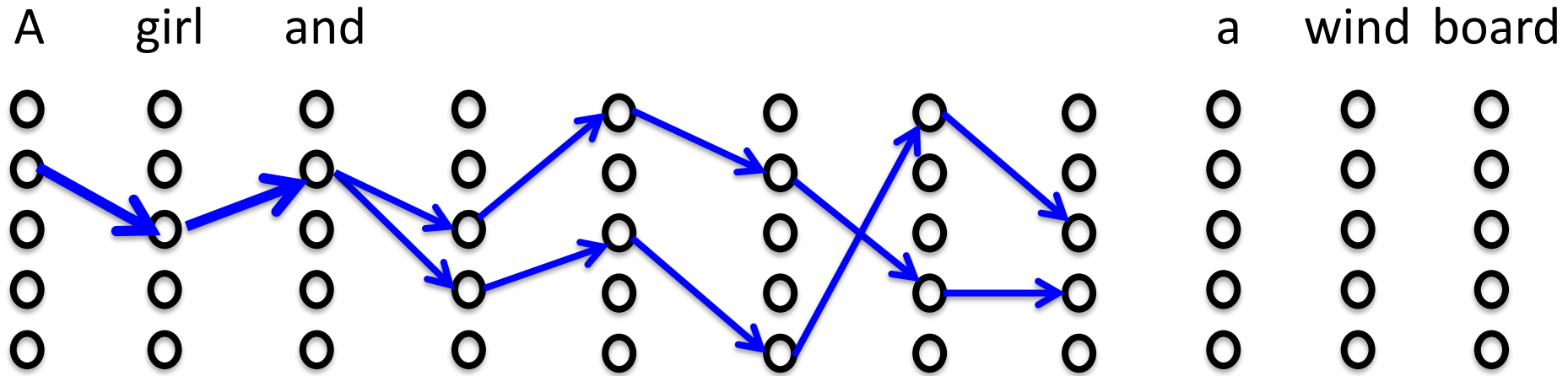
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



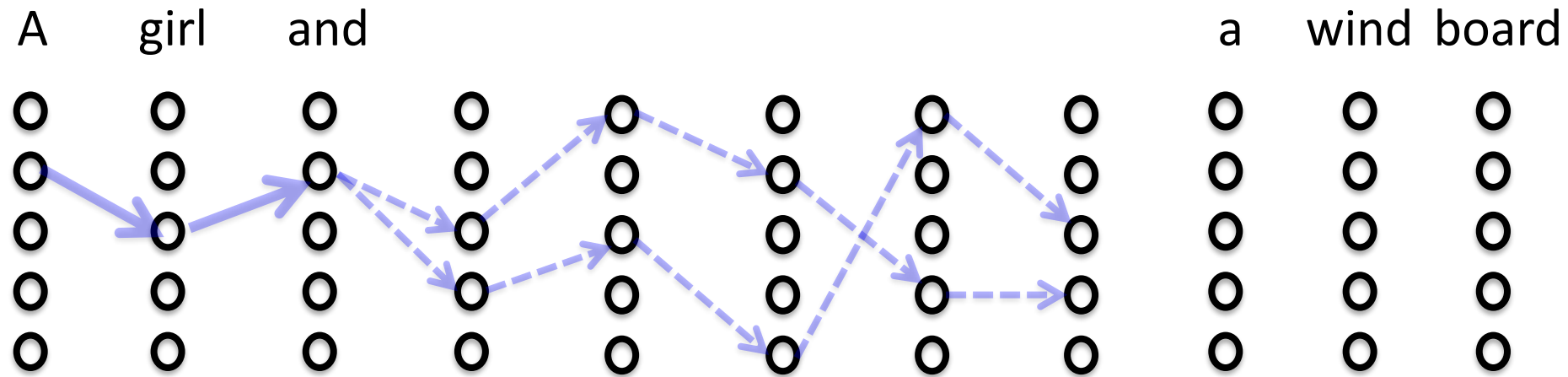
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



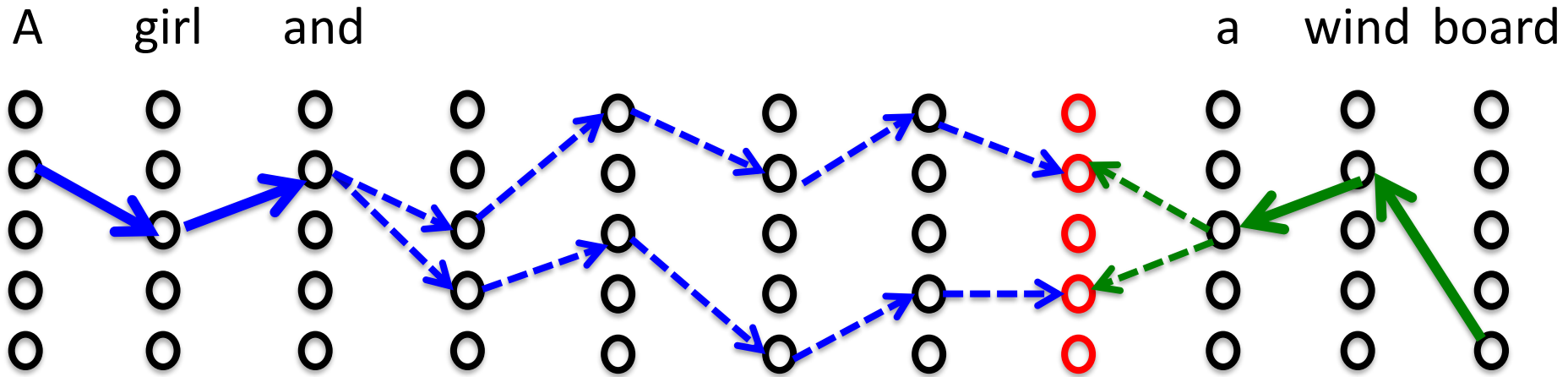
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



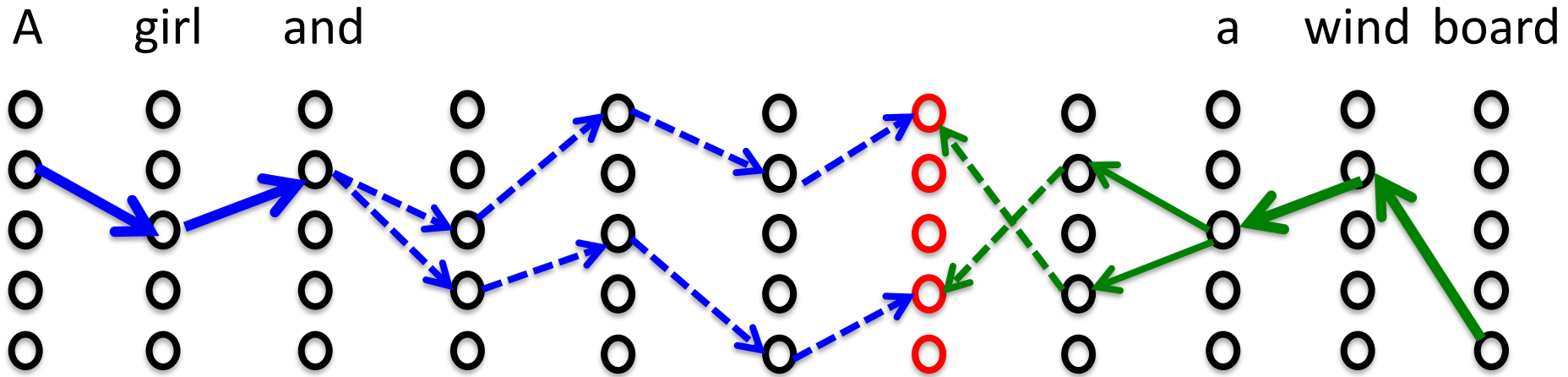
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



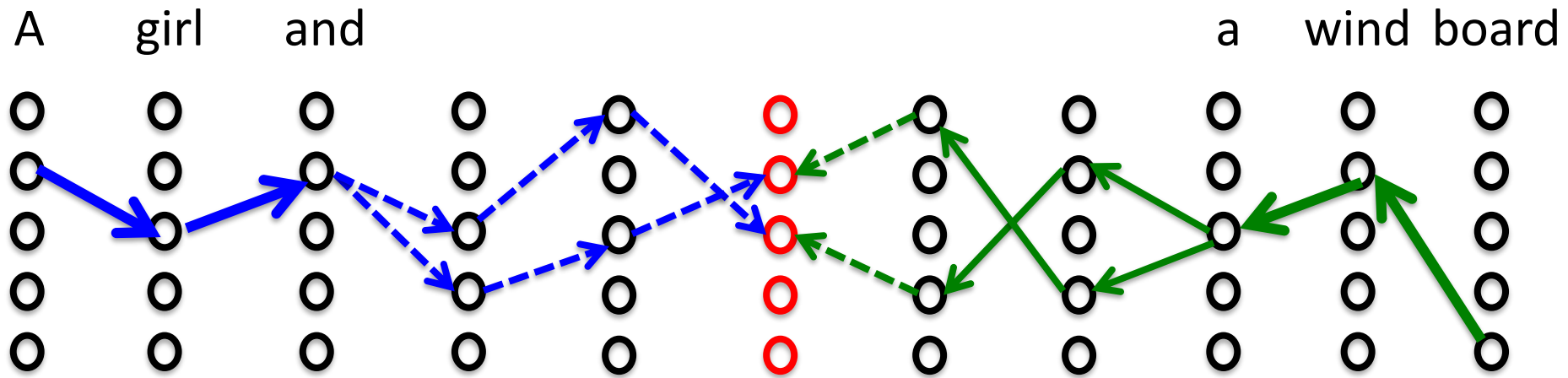
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



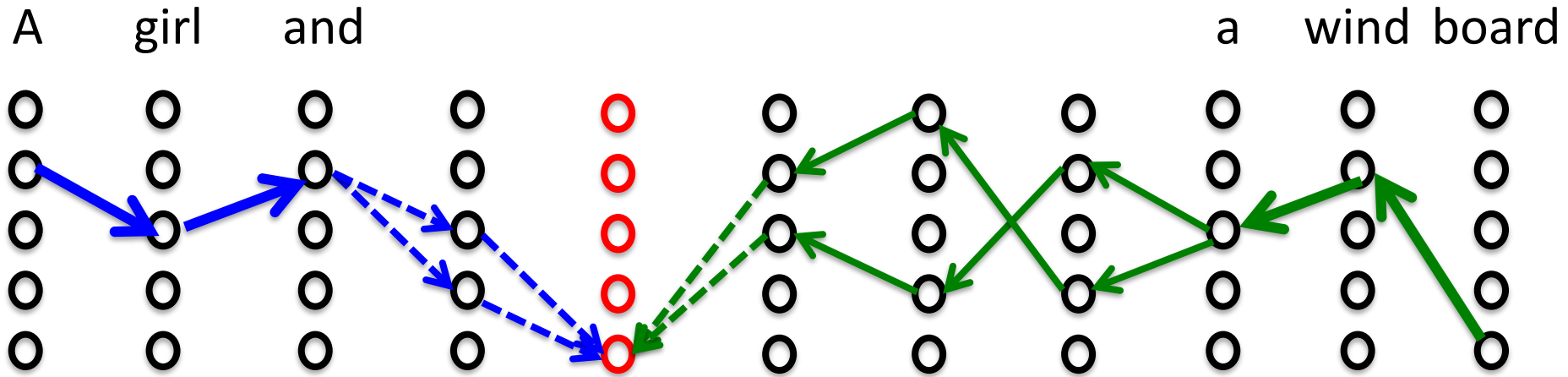
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



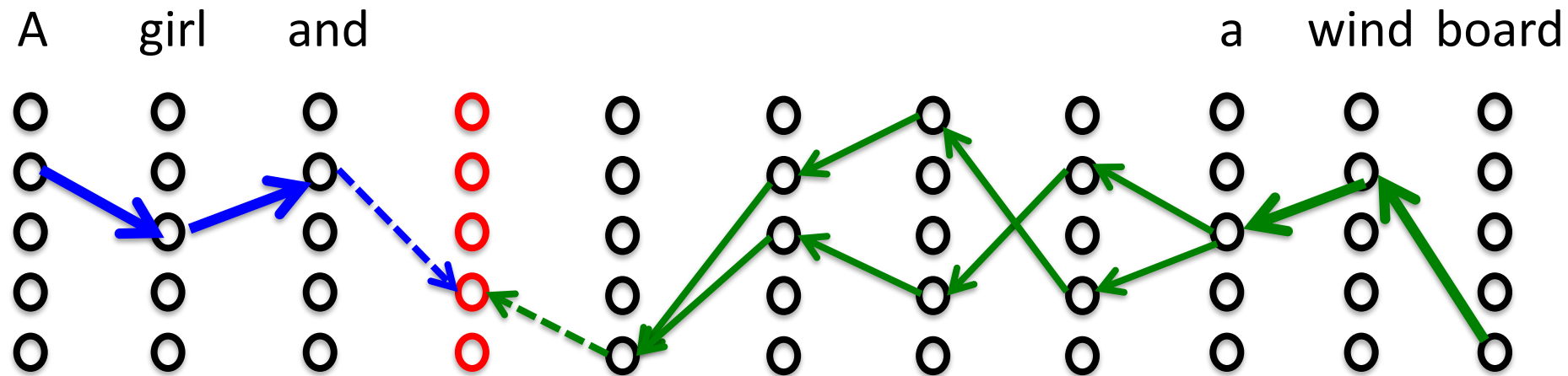
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



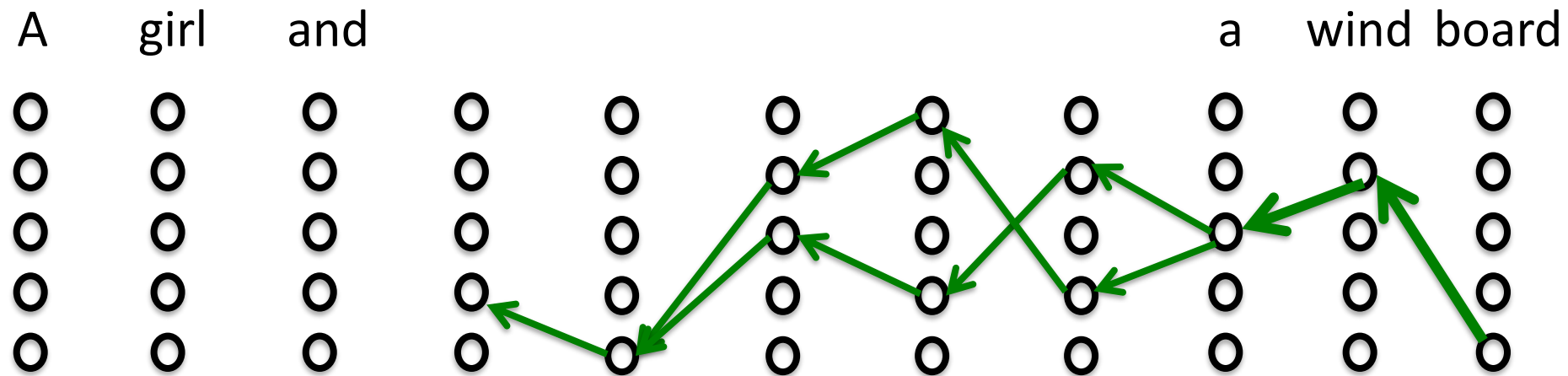
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



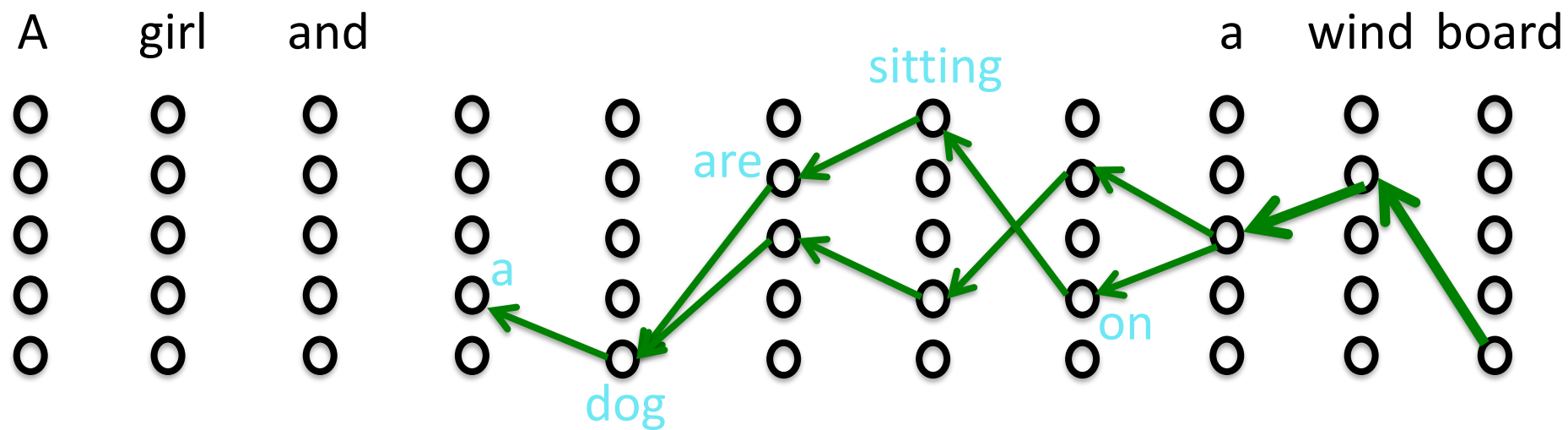
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



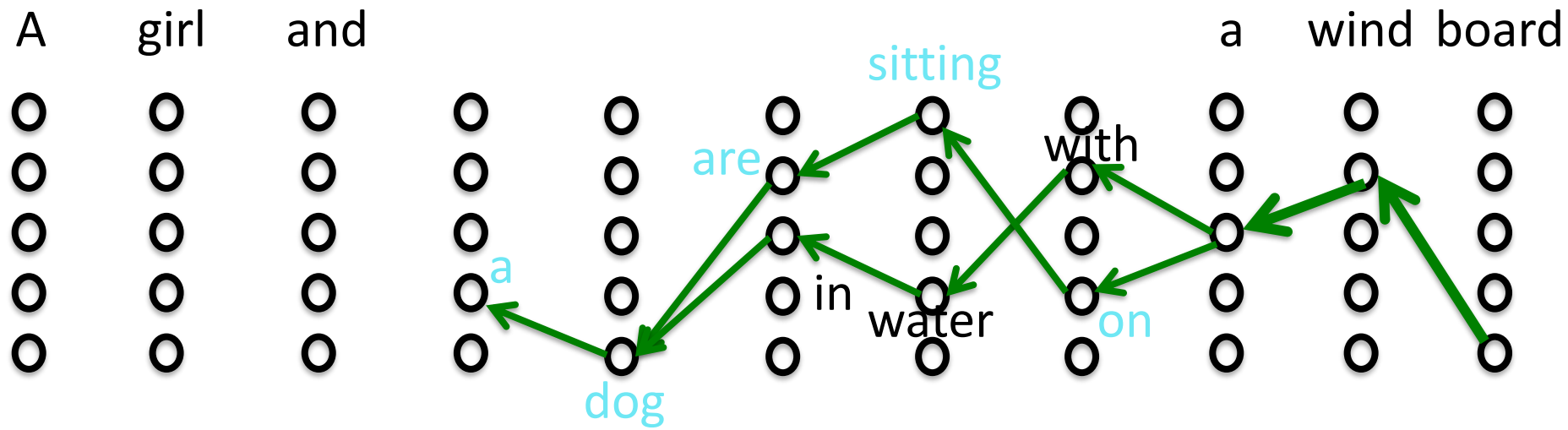
Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



Beam-Search Coordinate Descent (BSCD) in Bi-RNNs



Fill-in-the-Blank in Image captioning

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO
 - 123,000 images, 5 captions annotated by AMT

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO
 - 123,000 images, 5 captions annotated by AMT
 - Val: 5000, test: 5000, train: rest
 - consistent with Neuraltalk2

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO
 - 123,000 images, 5 captions annotated by AMT
 - Val: 5000, test: 5000, train: rest
 - consistent with Neuraltalk2
- Remove words in the middle

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO
 - 123,000 images, 5 captions annotated by AMT
 - Val: 5000, test: 5000, train: rest
 - consistent with Neuraltalk2
- Remove words in the middle

$r = 0.25$: A girl and a dog are balancing on a wind board

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO
 - 123,000 images, 5 captions annotated by AMT
 - Val: 5000, test: 5000, train: rest
 - consistent with Neuraltalk2
- Remove words in the middle

$r = 0.50$: A girl and a dog are balancing on a wind board

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO
 - 123,000 images, 5 captions annotated by AMT
 - Val: 5000, test: 5000, train: rest
 - consistent with Neuraltalk2
- Remove words in the middle

$r = 0.75$: A girl and a dog are balancing on a windboard

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO
 - 123,000 images, 5 captions annotated by AMT
 - Val: 5000, test: 5000, train: rest
 - consistent with Neuraltalk2
- Remove words in the middle

$r = 0.75$: A girl and a dog are balancing on a windboard
- Evaluation

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO
 - 123,000 images, 5 captions annotated by AMT
 - Val: 5000, test: 5000, train: rest
 - consistent with Neuraltalk2
- Remove words in the middle

$r = 0.75$: A girl and a dog are balancing on a wind board
- Evaluation
 - Full sentence BLEU, CIDEr, Meteor.

Fill-in-the-Blank in Image captioning

- Dataset: MSCOCO
 - 123,000 images, 5 captions annotated by AMT
 - Val: 5000, test: 5000, train: rest
 - consistent with Neuraltalk2
- Remove words in the middle

$r = 0.75$: A girl and a dog are balancing on a wind board
- Evaluation
 - Full sentence BLEU, CIDEr, Meteor.
 - Bad completion “A girl and sitting ”
 - > low n-gram match with humans

Qualitative Result



Qualitative Result



A person standing posing for a photo holding a glass of wine

Qualitative Result



URNN-f: A person standing in a room holding a cell glass of wine

Qualitative Result



URNN-b: A person standing a women is holding a glass of wine

Qualitative Result



BiRNN-BSCD: A person standing in a room while holding a glass of wine

Qualitative Result



Qualitative Result



A close up flowers and plants inside of a bowl

Qualitative Result



URNN-f: A close up of a vase with of a bowl

Qualitative Result



URNN-b A close vase that is sitting inside of a bowl

Qualitative Result



BiRNN-BSCD: A close up of flowers sitting inside of a bowl

Qualitative Result



Qualitative Result



A white hand holding a chocolate sprinkled donut

Qualitative Result



RNN-f: A white frosted doughnut with sprinkles sprinkled donut

Qualitative Result



URNN-b: A white close up of a sprinkled donut

Qualitative Result



BiRNN-BSCD: A white hand holding a pink sprinkled donut

Qualitative Result



Qualitative Result



A woman holding a pizza in her hand in the middle of a kitchen

Qualitative Result



URNN-f: A woman holding a plate of food in a kitchen middle of a kitchen

Qualitative Result



URNN-b: A woman holding a woman preparing food in the middle of the kitchen

Qualitative Result



BiRNN-BSCD: A woman holding a plate of food sitting in the middle of a kitchen

Qualitative Result



BiRNN-BSCD: A woman holding a plate of food sitting in the middle of a kitchen

Fill-in-the-Blank in Image captioning

(Length of Ground-truth is known)

	$r = 0.25$	$r = 0.5$	$r = 0.75$
URNN-f	6.628	3.915	2.034

Table 1. Comparison of different approaches on MSCOCO test dataset(metric: CIDEr; r : the fraction of removed words.)

Fill-in-the-Blank in Image captioning

(Length of Ground-truth is known)

	$r = 0.25$	$r = 0.5$	$r = 0.75$
URNN-f	6.628	3.915	2.034
URNN-b	6.639	3.965	2.532

Table 1. Comparison of different approaches on MSCOCO test dataset(metric: CIDEr; r : the fraction of removed words.)

Fill-in-the-Blank in Image captioning

(Length of Ground-truth is known)

	$r = 0.25$	$r = 0.5$	$r = 0.75$
URNN-f	6.628	3.915	2.034
URNN-b	6.639	3.965	2.532
URNN-fb-max	6.758	4.054	2.297

Table 1. Comparison of different approaches on MSCOCO test dataset(metric: CIDEr; r : the fraction of removed words.)

Fill-in-the-Blank in Image captioning

(Length of Ground-truth is known)

	$r = 0.25$	$r = 0.5$	$r = 0.75$
URNN-f	6.628	3.915	2.034
URNN-b	6.639	3.965	2.532
URNN-fb-max	6.758	4.054	2.297
URNN-fb-BiRNN	6.622	3.92	2.088

Table 1. Comparison of different approaches on MSCOCO test dataset(metric: CIDEr; r : the fraction of removed words.)

Fill-in-the-Blank in Image captioning

(Length of Ground-truth is known)




	$r = 0.25$	$r = 0.5$	$r = 0.75$
URNN-f	6.628	3.915	2.034
URNN-b	6.639	3.965	2.532
URNN-fb-max	6.758	4.054	2.297
URNN-fb-BiRNN	6.622	3.92	2.088
BiRNN-BSCD	7.262	4.413	2.534
	 7.4%	 8.8%	 7%

Table 1. Comparison of different approaches on MSCOCO test dataset(metric: CIDEr; r : the fraction of removed words.)

Visual Madlib

- 360,001 focused descriptions for 10,738 images
- Two evaluation tasks:
 - Multiple-choice question-answering
 - **Fill-in-the-blank** image description
- 12 types of fill-in-the-blanks:

Type 7: object's affordance



People could relax on the couches

Type 12: pair's relationship



Person B is putting food in the bowl

Visual Madlib

(Length of Ground-truth is known)

	Type 7		Type 12	
	Bleu-1	Bleu-2	Bleu-1	Bleu-2
nCCA(box)	0.6	0.11	0.48	0.08
URNN-f	0.315	0.140	0.275	0.158
URNN-b	0.461	0.285	0.346	0.212
URNN-fb-max	0.449	0.275	0.345	0.211
BiRNN-BSCD	0.470	0.300	0.353	0.231

Table 2. Comparison of different approaches on Madlibs test dataset

Visual Madlib

(Length of Ground-truth is known)



	Type 7		Type 12	
	Bleu-1	Bleu-2	Bleu-1	Bleu-2
nCCA(box)	0.6	0.11	0.48	0.08
URNN-f	0.315	0.140	0.275	0.158
URNN-b	0.461	0.285	0.346	0.212
URNN-fb-max	0.449	0.275	0.345	0.211
BiRNN-BSCD	0.470	0.300	0.353	0.231
		 5%		
			 8.9%	

Table 2. Comparison of different approaches on Madlibs test dataset

Conclusion

- Beam-based Top-B MAP Inference algorithm for Bi-RNNs
- Any Partial-MAP estimation in sequence prediction problem

Thank You !

Q&A

Fill-in-the-Blank in Image captioning

(Length of Ground-truth is unknown)

	$r = 0.25$	$r = 0.5$	$r = 0.75$
URNN-fb-max	6.971	4.203	2.665
BiRNN-BSMP	8.356	5.40	3.544

Table 2. Comparison of different approaches on MSCOCO test dataset(metric: CIDEr; r: the fraction of removed words.)