Microsoft Research

Mind's Eye: A Recurrent Visual Representation for Image Caption Generation

Xinlei Chen[1], Larry Zitnick[2]

[1]Carnegie Mellon University, [2]Microsoft Research

Carnegie Mellon University

Overview

Goal: Learn a bi-directional mapping between images and their sentence-based descriptions

Usage:

- Bi-directional retrieval
- Caption generation

Key Motivation:

- Visual representations help build long-term memory
- A good caption should capture and help reconstruct the visual representation.

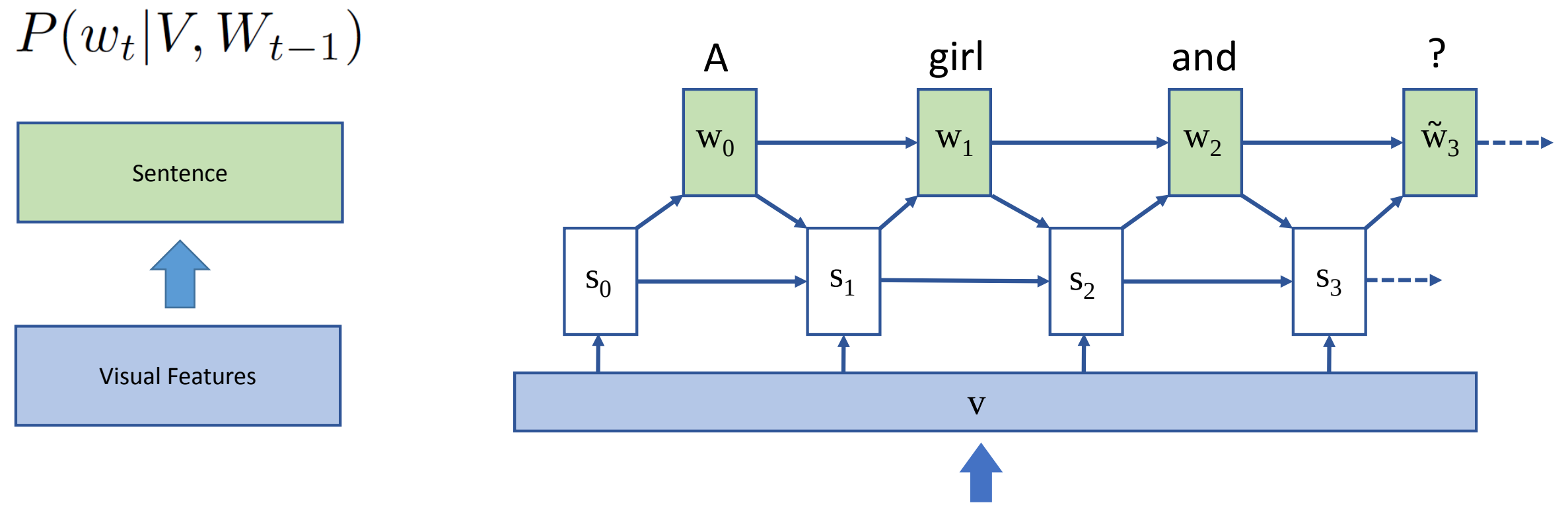
Evolving visual memory…

A girl and boy knocked down a tower.

Background

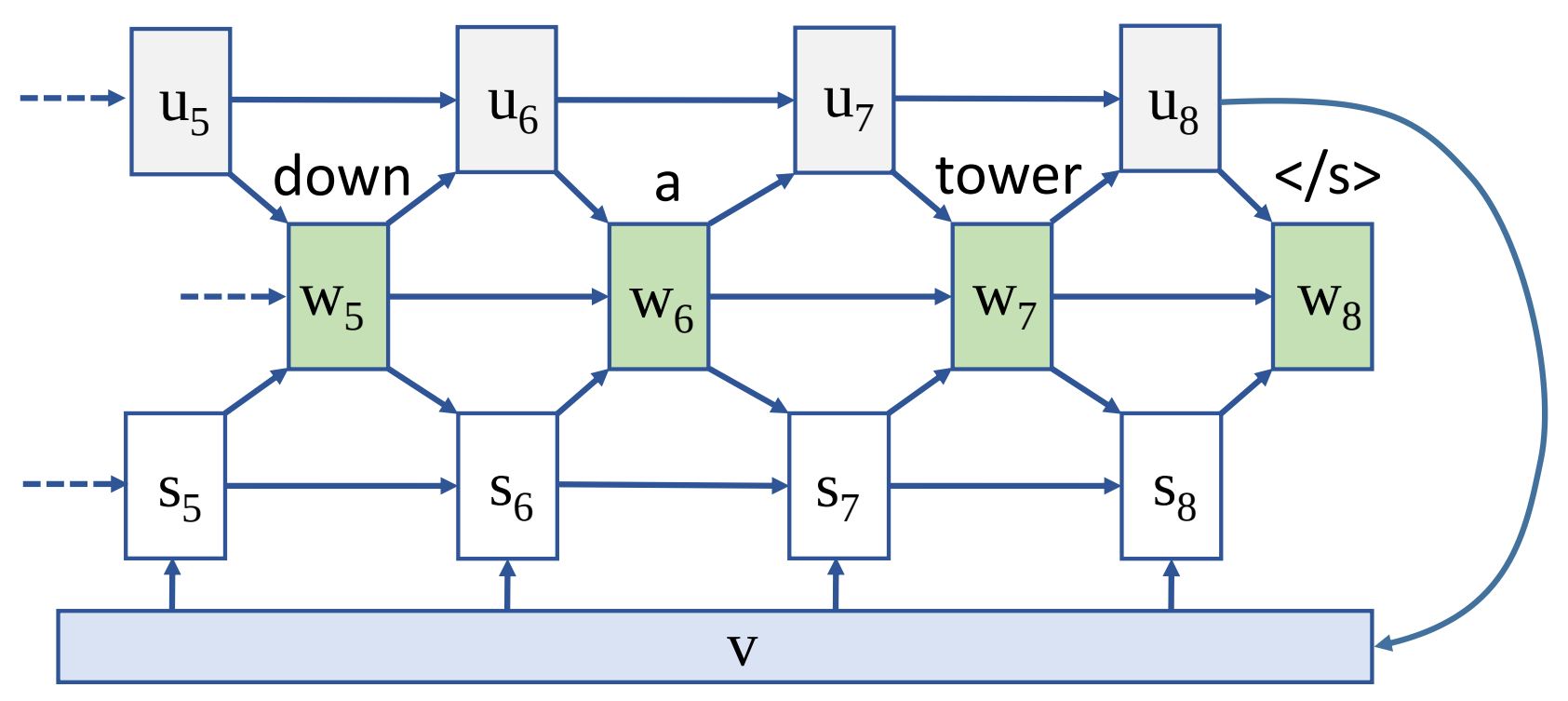
Previous RNN Model:

$P(w_t|V,W_{t-1})$



Context dependent recurrent neural network language model. T. Mikolov and G. Zweig, SLT 2012.

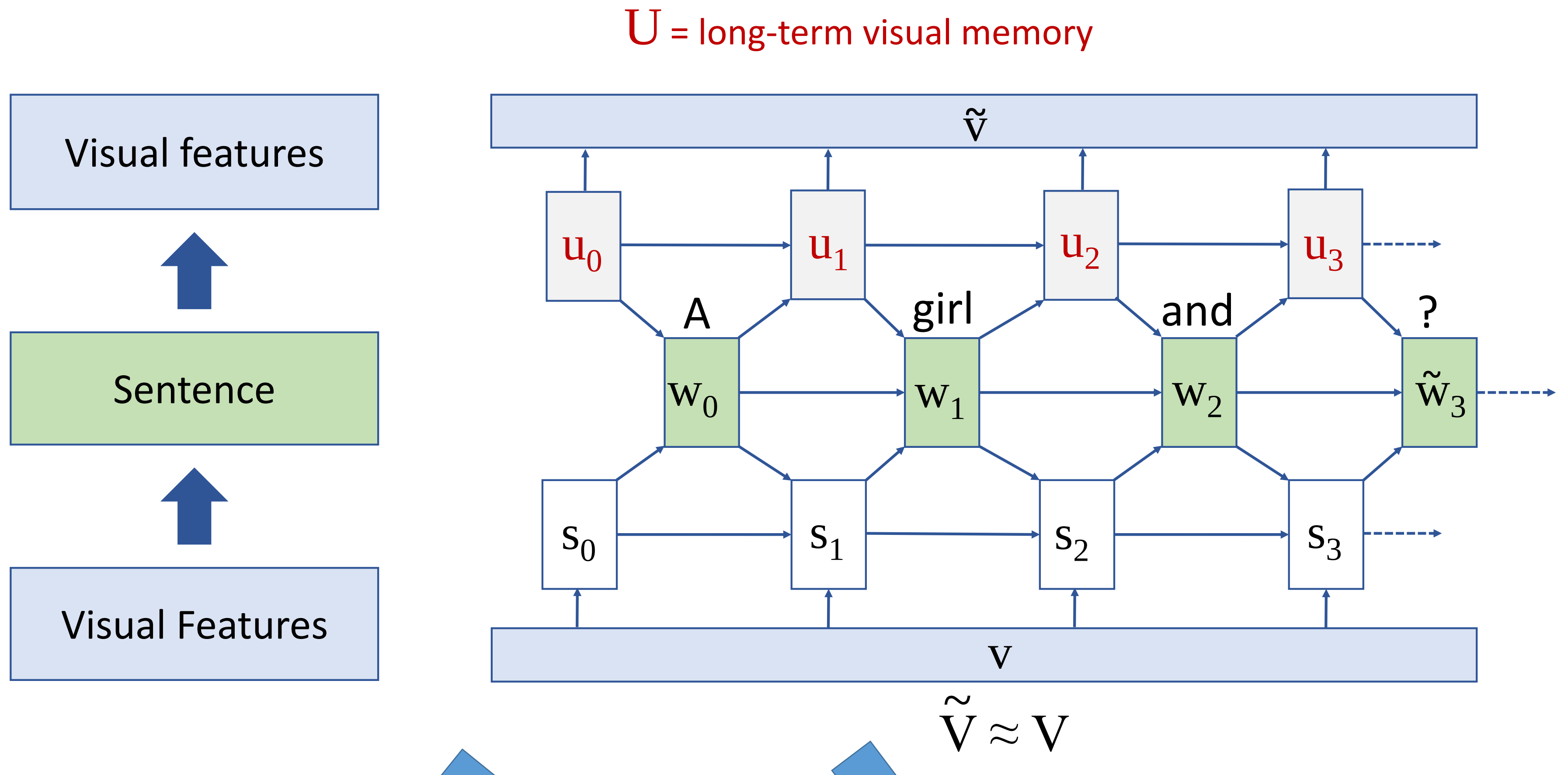
First Attempt:



Long Term Memory

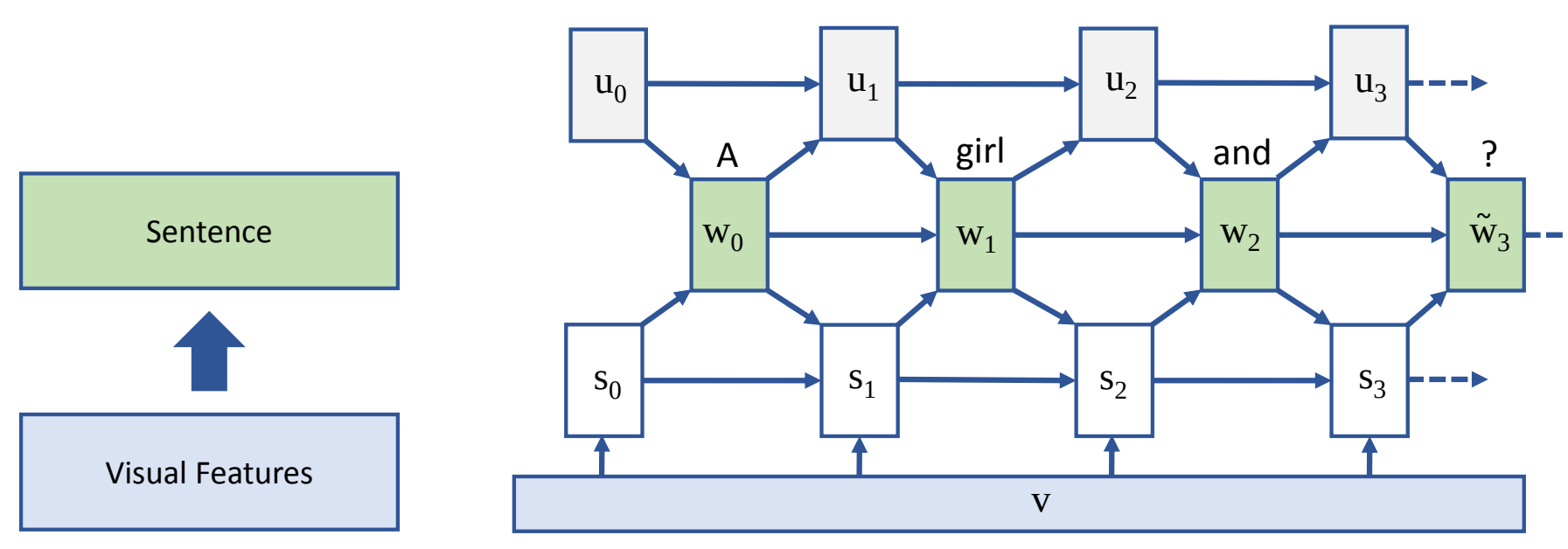
Our Model

$$P(w_t,V|W_{t-1}) = P(w_t|V,W_{t-1})\,P(V|W_{t-1})$$



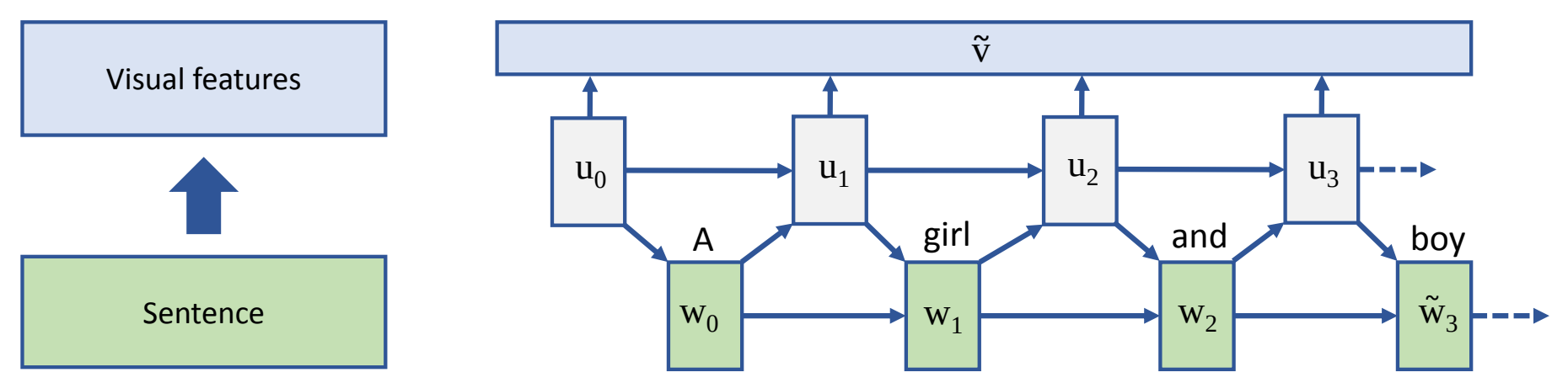
$$\tilde{V} \approx V$$

Caption Generation



$P(w_t|V,W_{t-1})$

Visual Feature Reconstruction



$P(V|W_{t-1})$

Training:

- Per stage model, every step tries to reconstruct the image
- Weight update from visual memory to image is performed from end to start

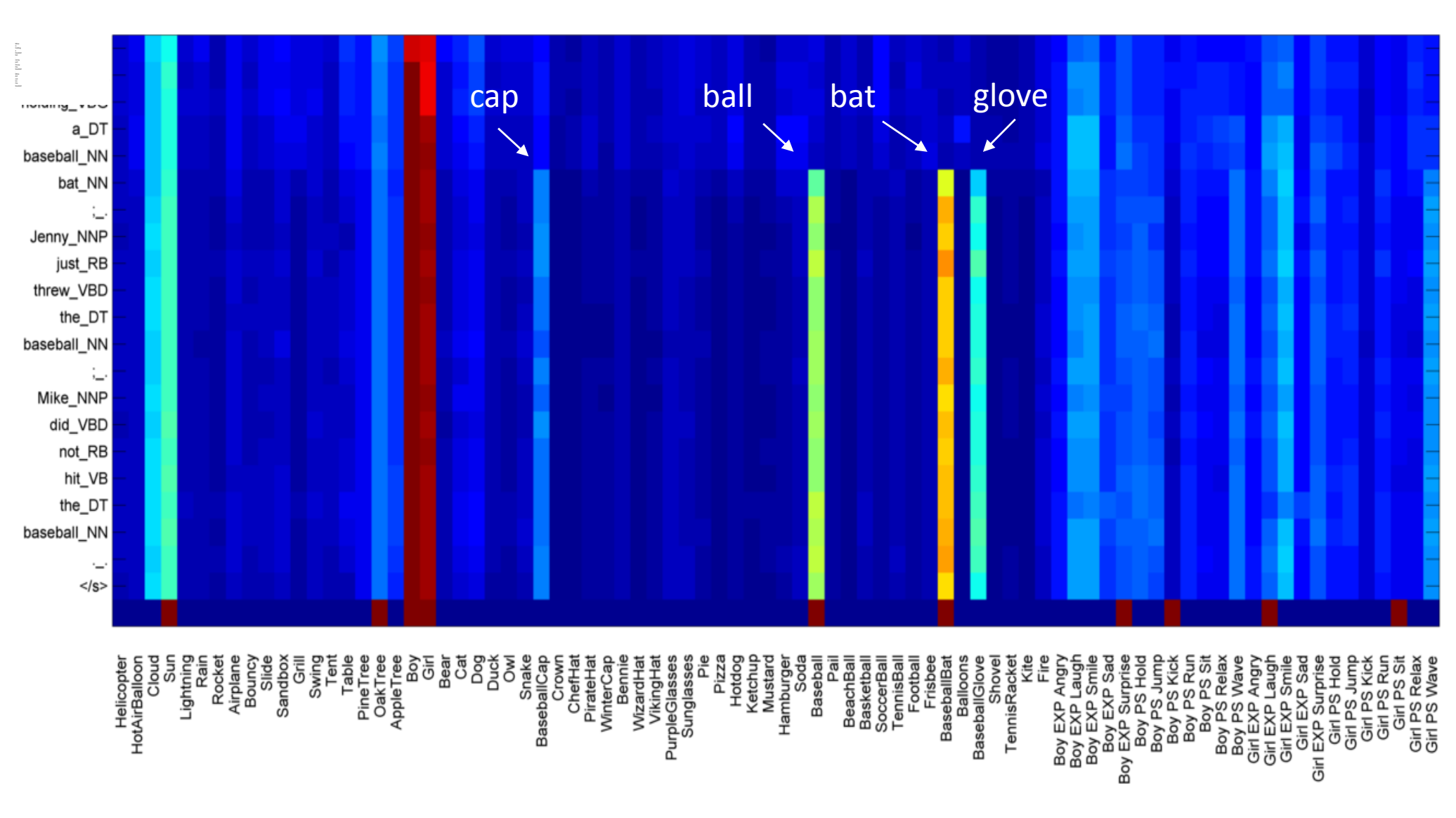
Retrieval:

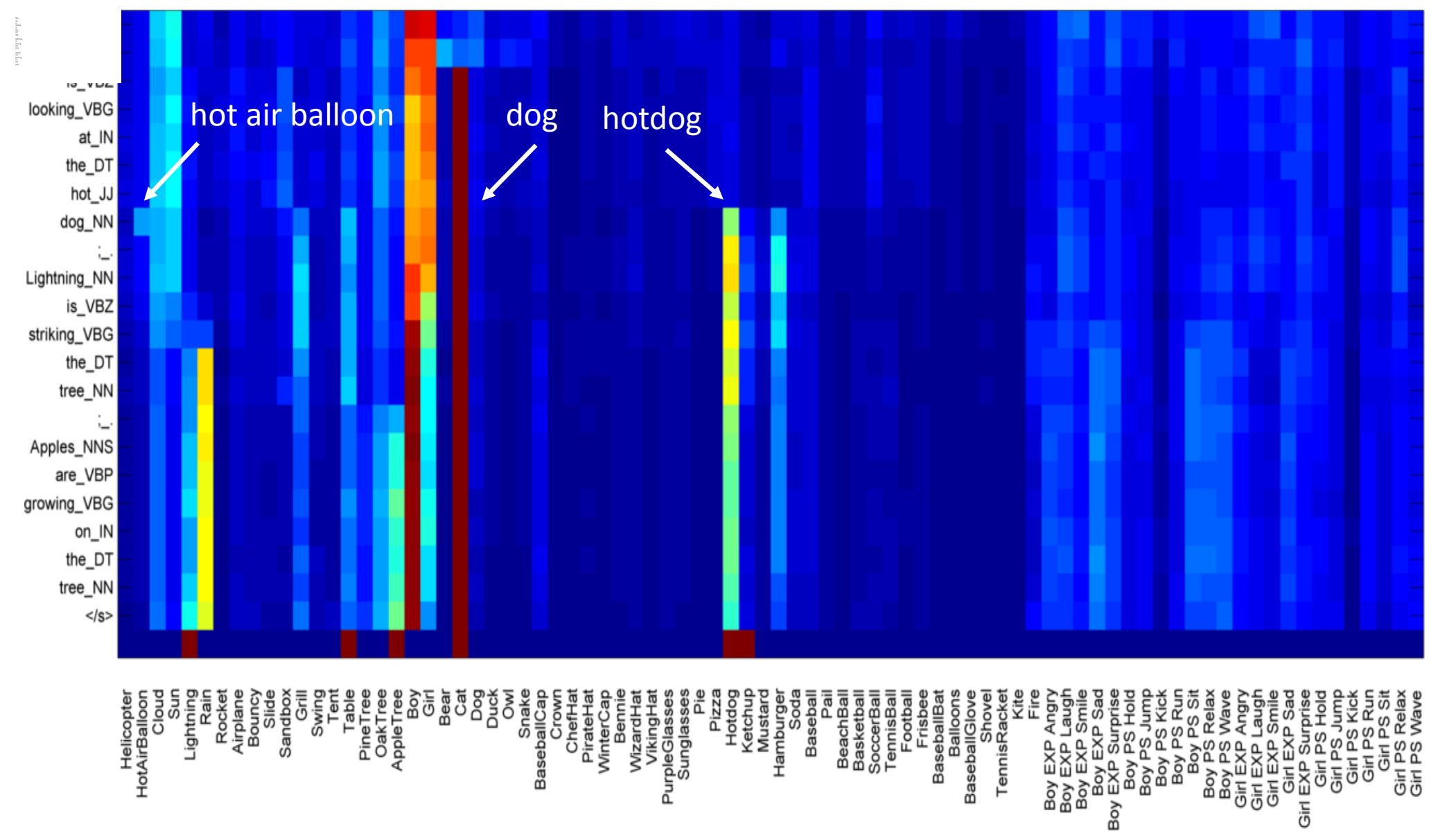
- Given a sentence, evaluate the likelihood that it can be generated by using each image as an input
- Image to sentence retrieval is normalized by sentence length
- Using visual memory helps the performance

Generation:

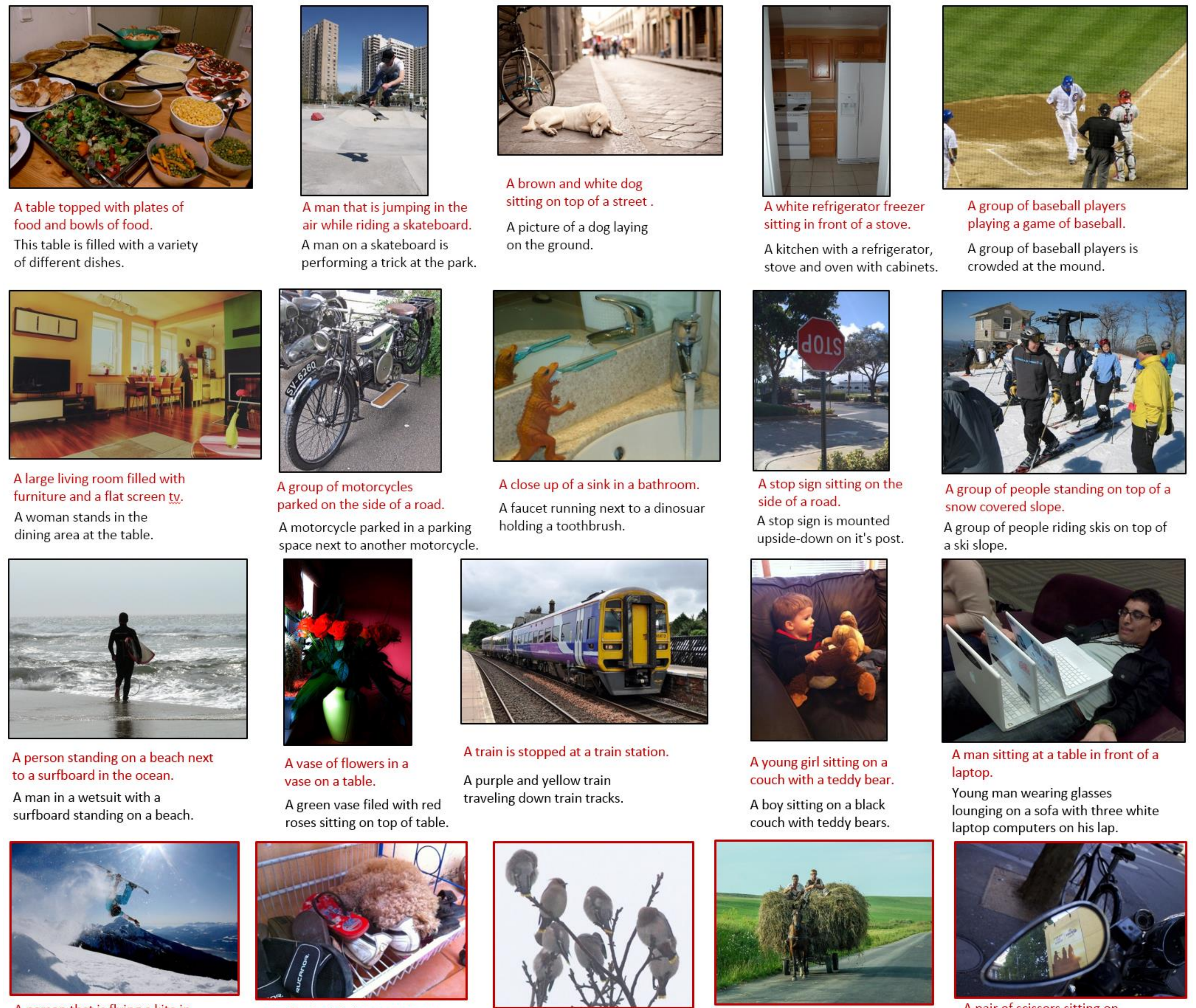
- First sample sentence length from a prior
- With fixed length, sample the most likely caption

Visual Feature Reconstruction





Results



A table topped with plates of food and bowls of food.
This table is filled with a variety of different dishes.

A man that is jumping in the air while riding a skateboard.
A man on a skateboard is performing a trick at the park.

A brown and white dog sitting on top of a street .
A picture of a dog laying on the ground.

A white refrigerator freezer sitting in front of a stove.
A kitchen with a refrigerator, stove and oven with cabinets.

A group of baseball players playing a game of baseball.
A group of baseball players is crowded at the mound.

A large living room filled with furniture and a flat screen tv.
A woman stands in the dining area at the table.

A group of motorcycles parked on the side of a road.
A motorcycle parked in a parking space next to another motorcycle.

A close up of a sink in a bathroom.
A faucet running next to a dinosuar holding a toothbrush.

A stop sign sitting on the side of a road.
A stop sign is mounted upside-down on it's post.

A group of people standing on top of a snow covered slope.
A group of people riding skis on top of a ski slope.

A person standing on a beach next to a surfboard in the ocean.
A man in a wetsuit with a surfboard standing on a beach.

A vase of flowers in a vase on a table.
A green vase filed with red roses sitting on top of table.

A train is stopped at a train station.
A purple and yellow train traveling down train tracks.

A young girl sitting on a couch with a teddy bear.
A boy sitting on a black couch with teddy bears.

A man sitting at a table in front of a laptop.
Young man wearing glasses lounging on a sofa with three white laptop computers on his lap.

A person that is flying a kite in the snow .
A person up in the air, upside down while outside.

A stuffed teddy bear sitting on top of a piece of luggage.
This wire metal rack holds several pairs of shoes and sandals.

A group of people that are standing in front of a building.
Birds perch on a bunch of twigs in the winter.

A man riding a skateboard on top of a green field.
A horse drawn carriage traveling down the middle of a road.

A pair of scissors sitting on top of a white car.
A rear view mirror on a bike is reflecting a poster on building.

	Flickr 8K			Flickr 30K			MS COCO Val			MS COCO Test		
	PPL	BLEU	METEOR	PPL	BLEU	METEOR	PPL	BLEU	METEOR	BLEU	METEOR	CIDEr
RNN	17.5	4.5	10.3	23.0	6.3	10.7	16.9	4.7	9.8	-	-	-
RNN+IF	16.5	11.9	16.2	20.8	11.3	14.3	13.3	16.3	17.7	-	-	-
RNN+IF+FT	16.0	12.0	16.3	20.5	11.6	14.6	12.9	17.0	18.0	-	-	-
RNN+VGG	15.2	12.4	16.7	20.0	11.9	15.0	12.6	18.4	19.3	18.0	19.1	51.5
Our Approach	16.1	12.2	16.6	20.0	11.3	14.6	12.6	16.3	17.8	-	-	-
Our Approach+FT	15.8	12.4	16.7	19.5	11.6	14.7	12.0	16.8	18.1	16.5	18.0	44.8
Our Approach+VGG	15.1	13.1	16.9	19.1	12.0	15.2	11.6	18.8	19.6	18.4	19.5	53.1
Human	-	20.6	25.5	-	18.9	22.9	-	19.2	24.1	21.7	25.2	85.4

	Sentence Retrieval				Image Retrieval			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random Ranking	0.1	0.6	1.1	631	0.1	0.5	1.0	500
SDT-RNN [38]	4.5	18.0	28.6	32	6.1	18.5	29.0	29
DeViSE [12]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
DeepFE [20]	12.6	32.9	44.0	14	9.7	29.6	42.5	15
DeepFE+DECAF [20]	5.9	19.2	27.3	34	5.2	17.6	26.5	32
RNN+VGG	8.9	25.7	38.7	20.5	6.5	17.3	28.4	25
Our Approach (T)	9.6	29.1	41.6	17	7.0	23.6	33.6	23
Our Approach (T+I)	9.9	29.2	42.4	16	7.3	24.6	36.0	20
[16]	8.3	21.6	30.3	34	7.6	20.7	30.1	38
RNN+VGG	7.7	23.0	37.2	21	6.8	24.0	33.9	23.5
Our Approach (T)	8.1	24.4	39.1	19	7.4	25.0	37.5	21
Our Approach (T+I)	8.6	25.9	40.1	17	7.6	24.9	37.8	20
M-RNN [28]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
RNN+VGG	14.4	37.9	48.2	10	15.6	38.4	50.6	10
Our Approach (T)	15.2	39.8	49.3	8.5	16.4	40.9	54.8	9
Our Approach (T+I)	15.4	40.6	50.1	8	17.3	42.5	57.4	7

	PASCAL		
	PPL	BLEU	METEOR
Midge [33]	-	2.9	8.8
Baby Talk [24]	-	0.5	9.7
Our Approach	25.3	9.8	16.0
Our Approach+FT	24.6	10.4	16.3
Our Approach+VGG	23.8	12.0	17.6
Human	-	20.1	25.0

- **Human Evaluation**:
  - 5.1% of our captions (Our Approach + VGG) are preferred to human captions, and 15.9% of equal quality

Conclusions

- Explicit visual memory is helpful
- Visual memory can be learned even with a single image per sentence
- Simple RNNs can remember long-term concepts
- Model is decomposable for bi-directional generation