Learning a Recurrent Visual Representation for Image Caption Generation

Xinlei Chen and C. Lawrence Zitnick

Presented by Yu-Cheng Lin

CVPR 2015

Most slides from Xinlei Chen

From Labels to Texts

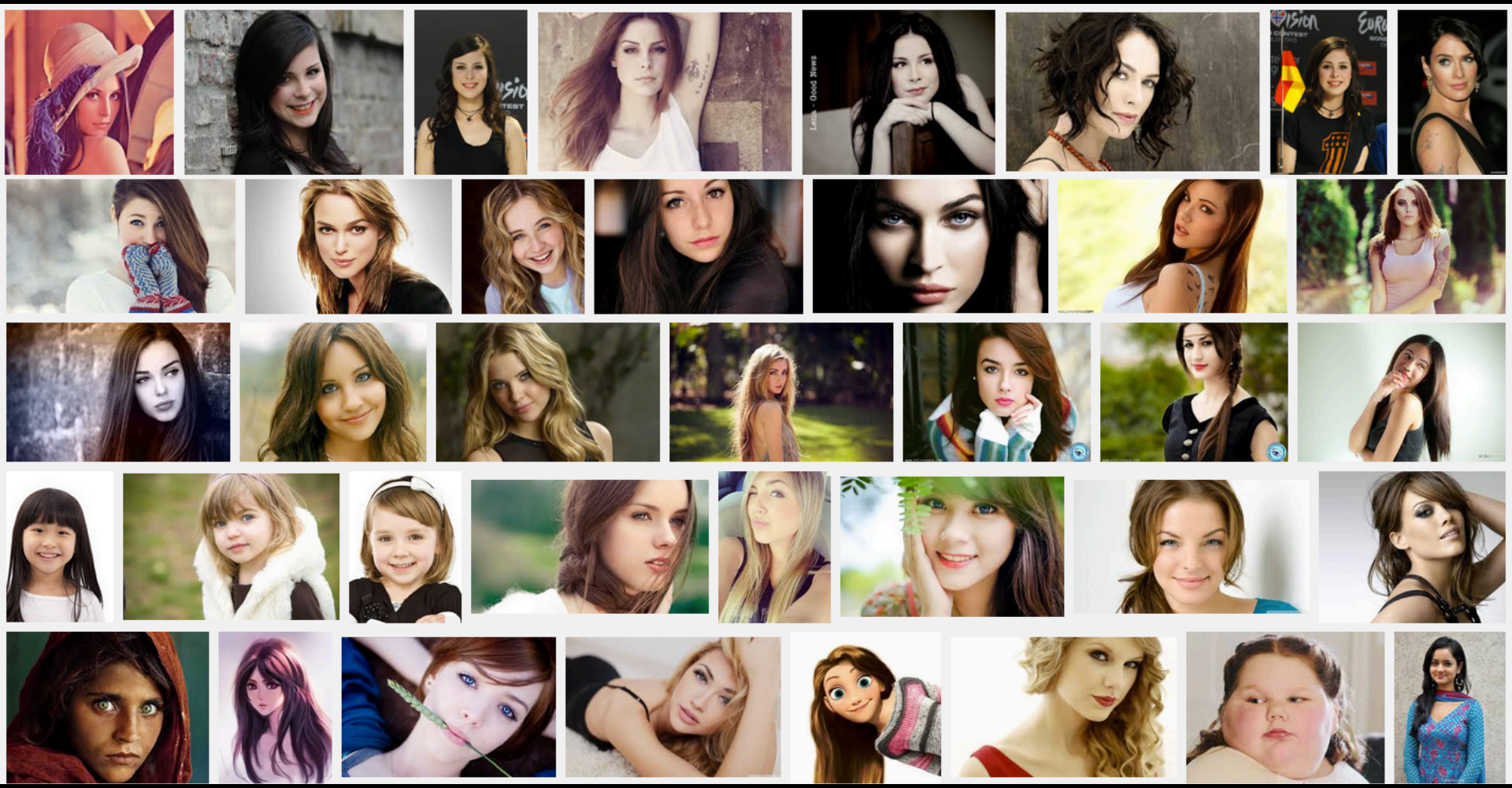
airplane

runway

From PASCAL VOC Challenge

TIME FOR HALLUCINATION!

A girl



A girl **sits**

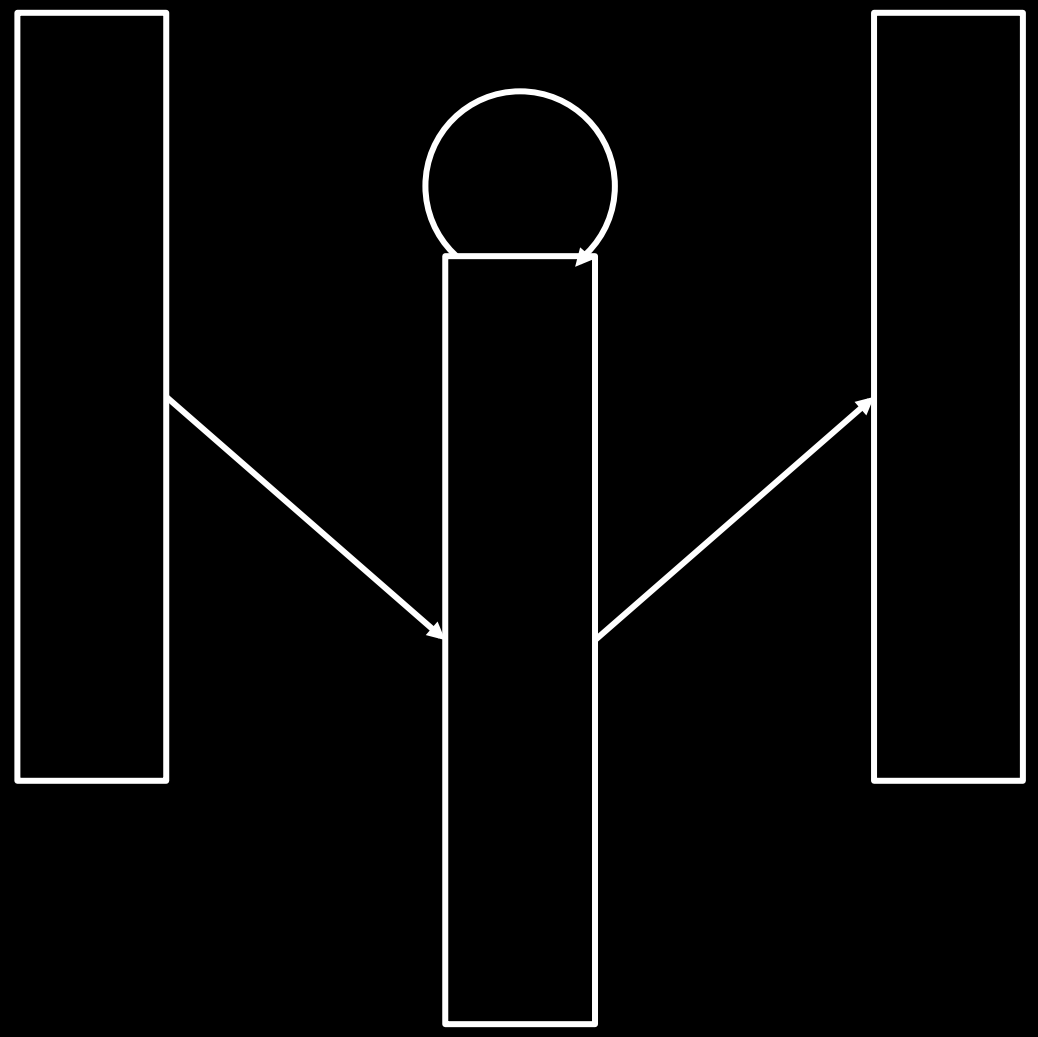
A girl sits on a tire **swing**

A girl sits on a tire swing **with a dog**.

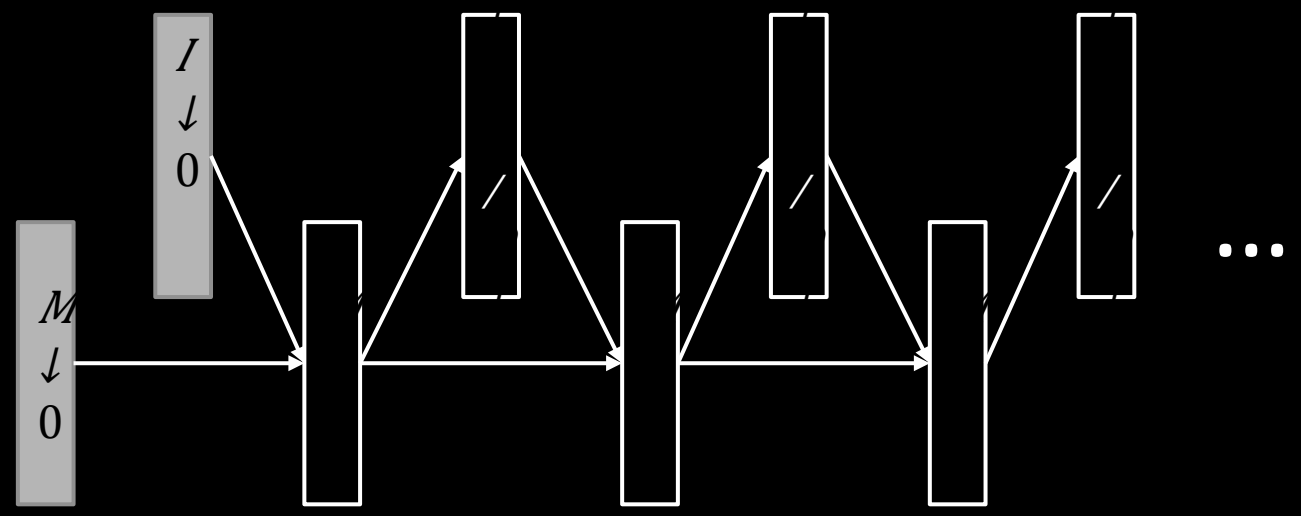
Can my computer learn something like that?

Model the sequence: memory

A simple model that memorizes …

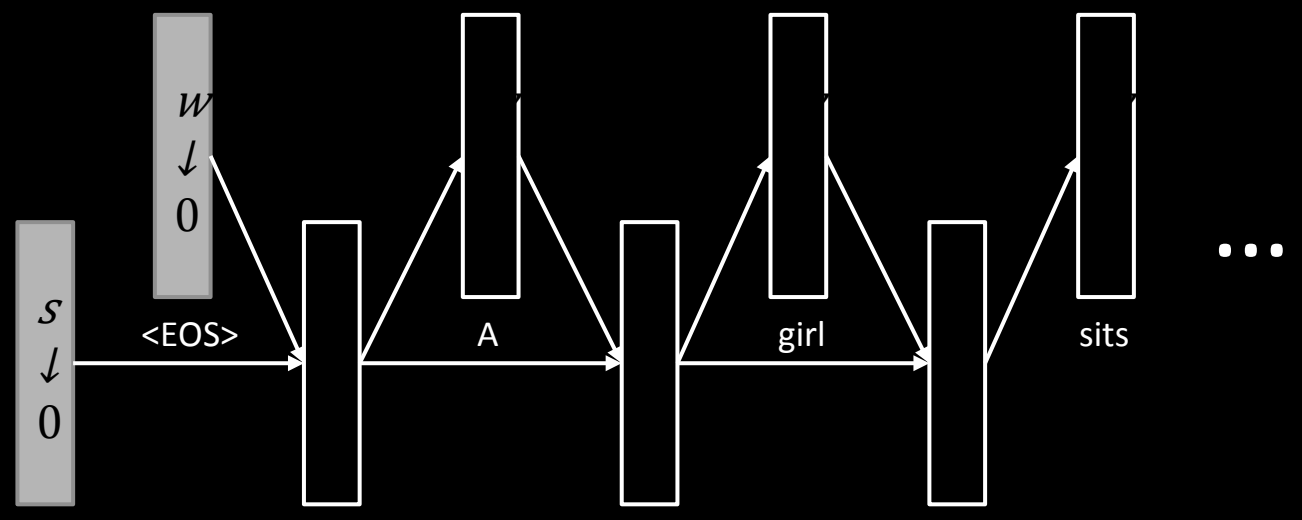


Un-roll in Time: Recurrent



Predicting the next

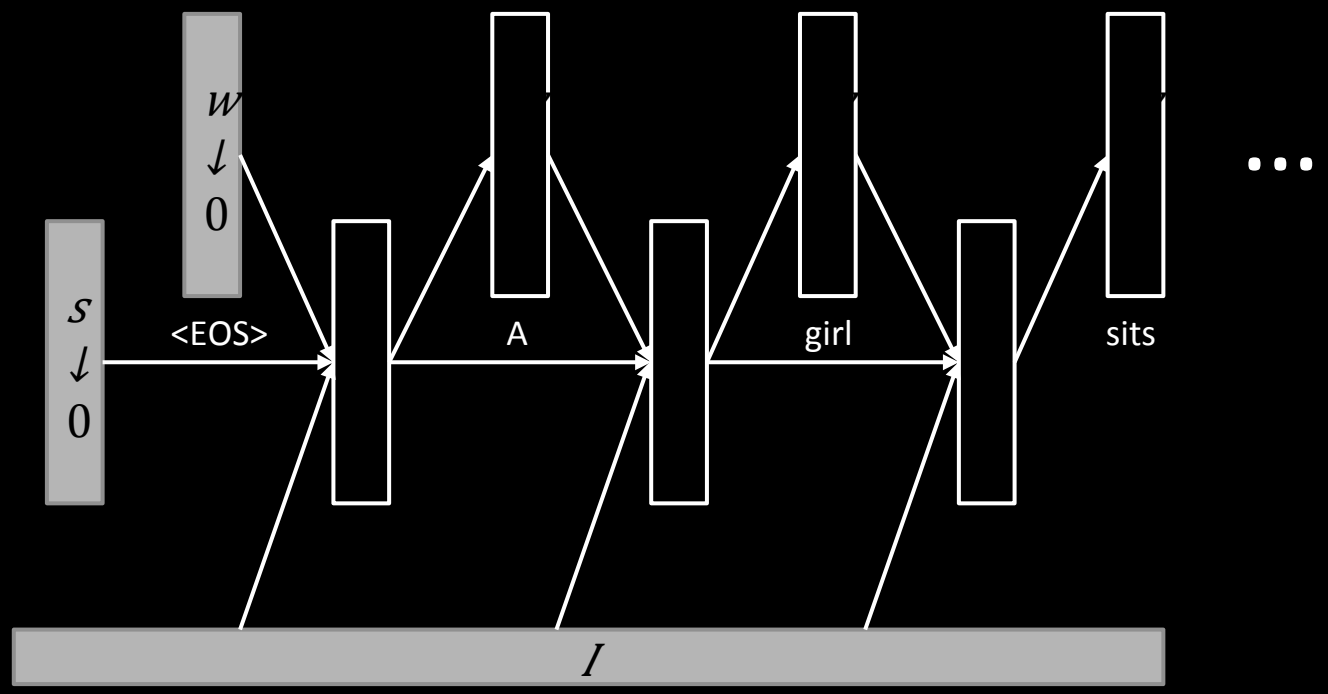
- Recurrent Neural Network based Language Model (RNNLM) [Mikolov et.al. 2010]



Now, back to our task

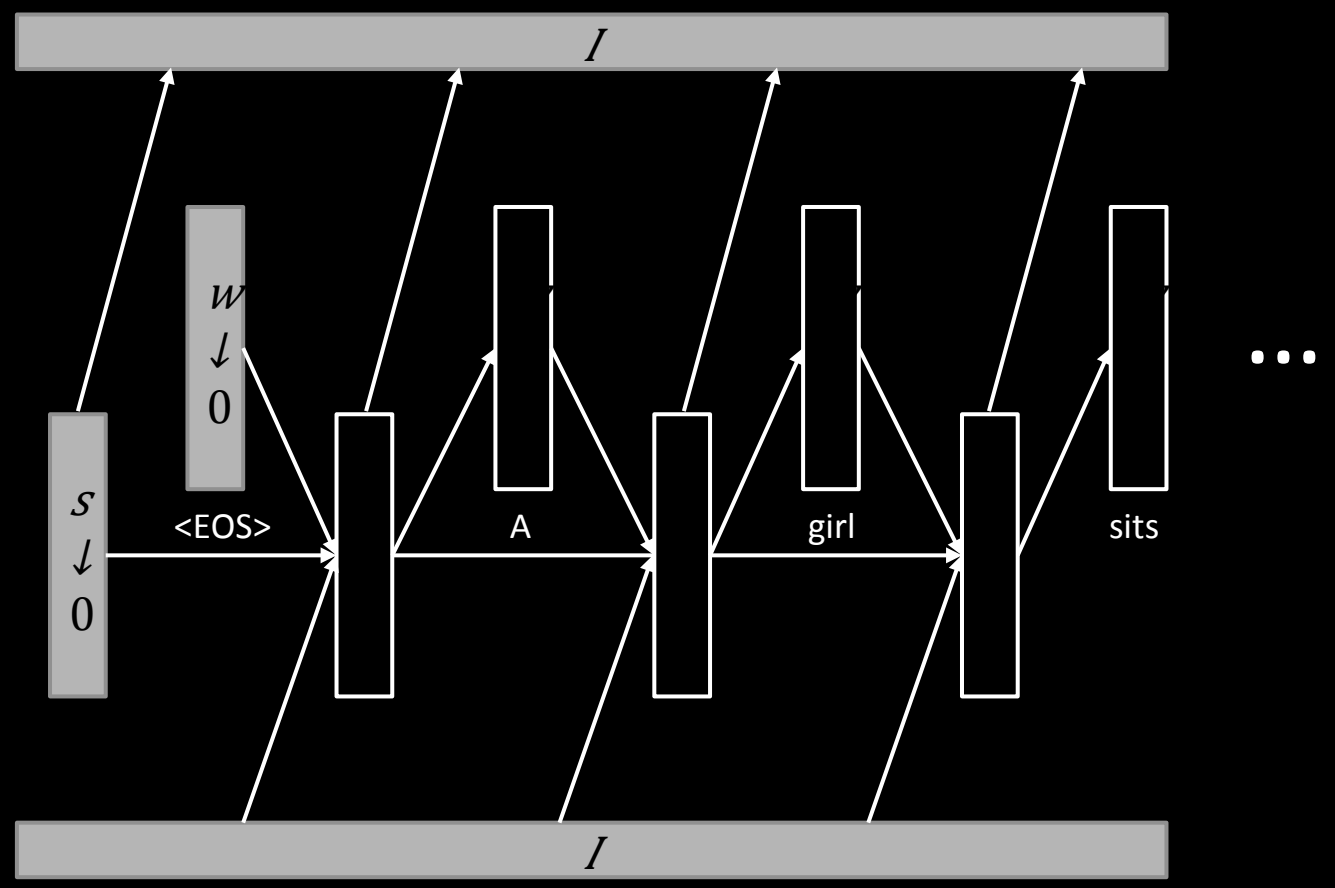
- Learn a model between images and sentences, so that:
  - Given an image, it can retrieve/**generate** a description
  - Given a description, it can retrieve/**hallucinate** an image

Generating Image Descriptions

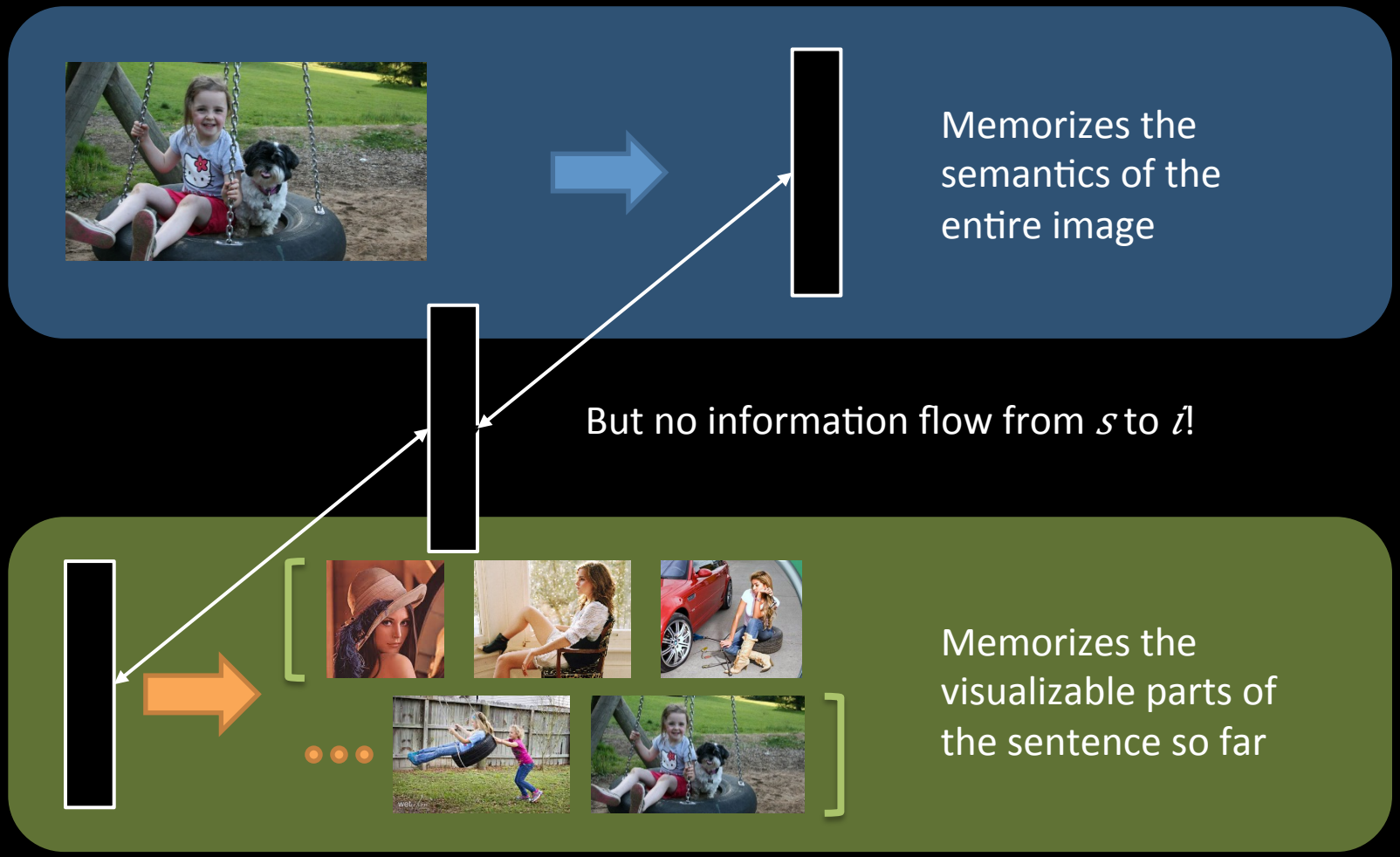


- EASY, Objective Function: $\Pi_t [P(w_t | I, H_t)]$

Provide more supervision!

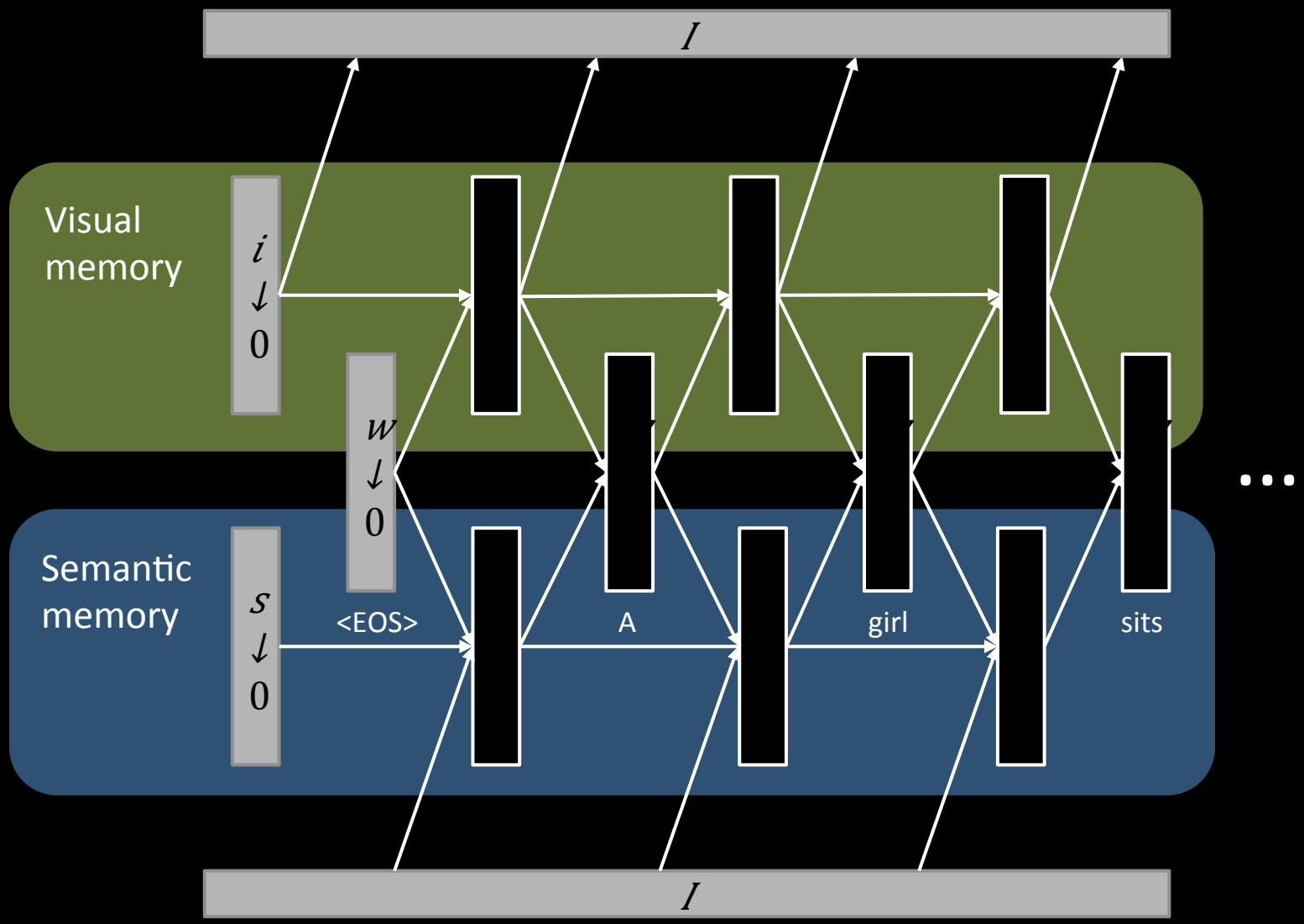


Solution: Split into 2!



Our Model (Per Stage)

$\Pi_{\downarrow t} [P(w_{\downarrow t}, I | H_{\downarrow t})]$



Sentence generation

$\Pi_t [P(w_t | I, H_t)]$

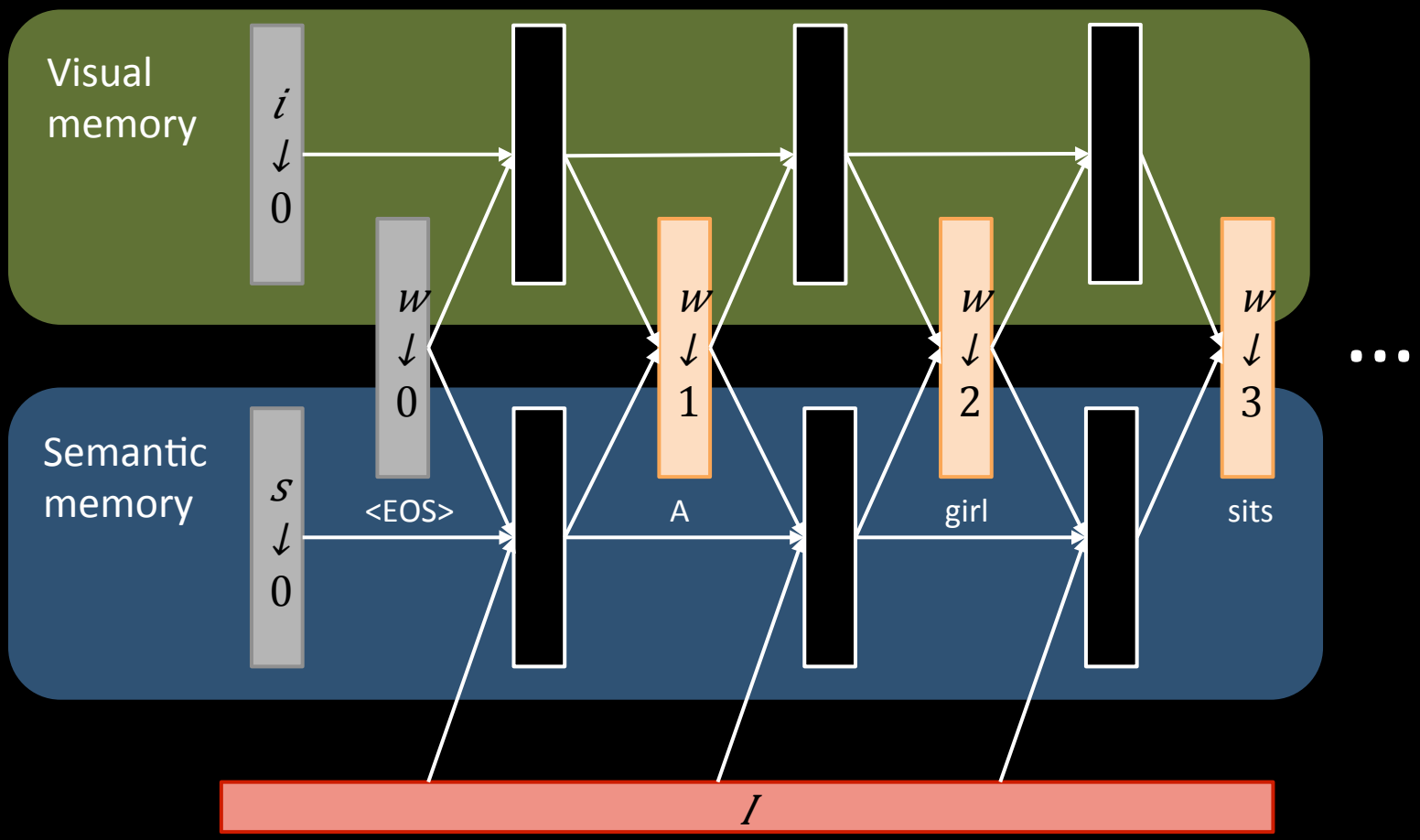
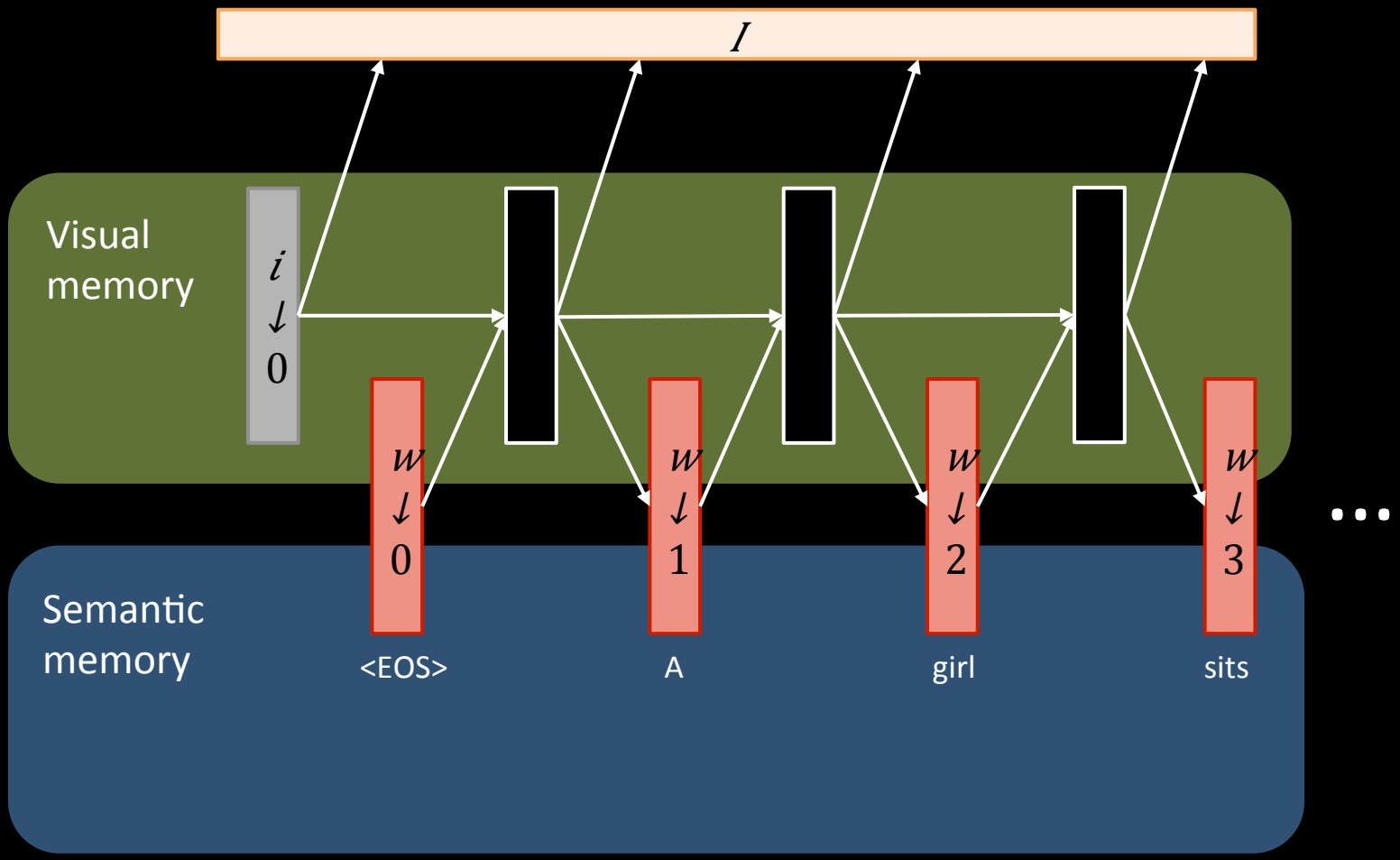


Image generation

$\Pi_{\downarrow t} [P(I|H_{\downarrow t})]$



Summary: Objective Function

- Language:

$$\pi_t[P(w_t|I, H_t)]$$

- Image:

$$\pi_t[P(I|H_t)]$$

- Together:

$$\pi_t[P(I, w_t|H_t)]$$

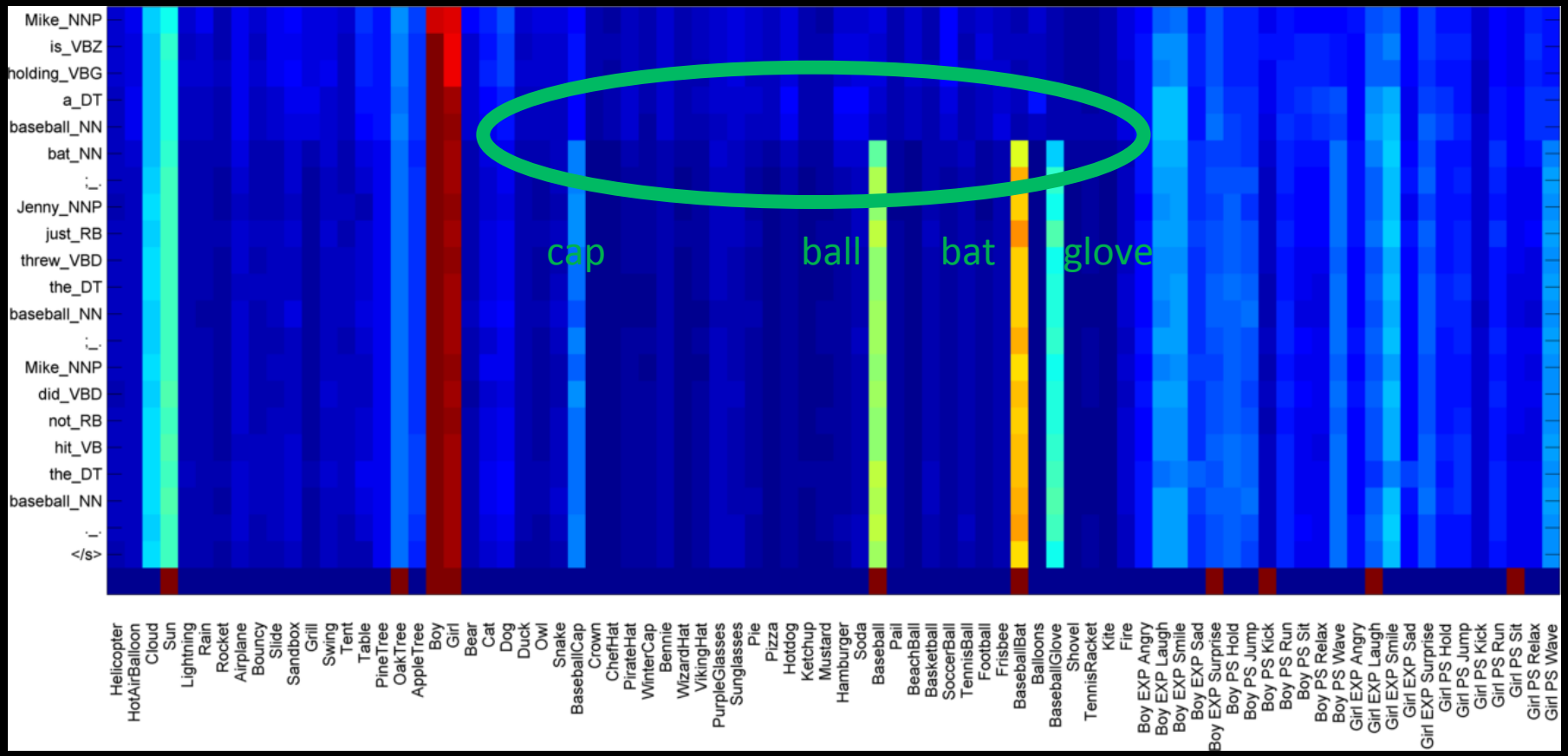
How does it work?

- Test-bed: Clip-art Dataset

  - Jenny is catching the ball;
  - Mike is kicking the ball;
  - The table is next to the tree.

  - Sentence (Vertical): concatenation of the 3 simple sentences
  - Image (Horizontal): 82 binary variable indicating whether the clip-art/attribute occurs

Image Feature Generation (Baseball)

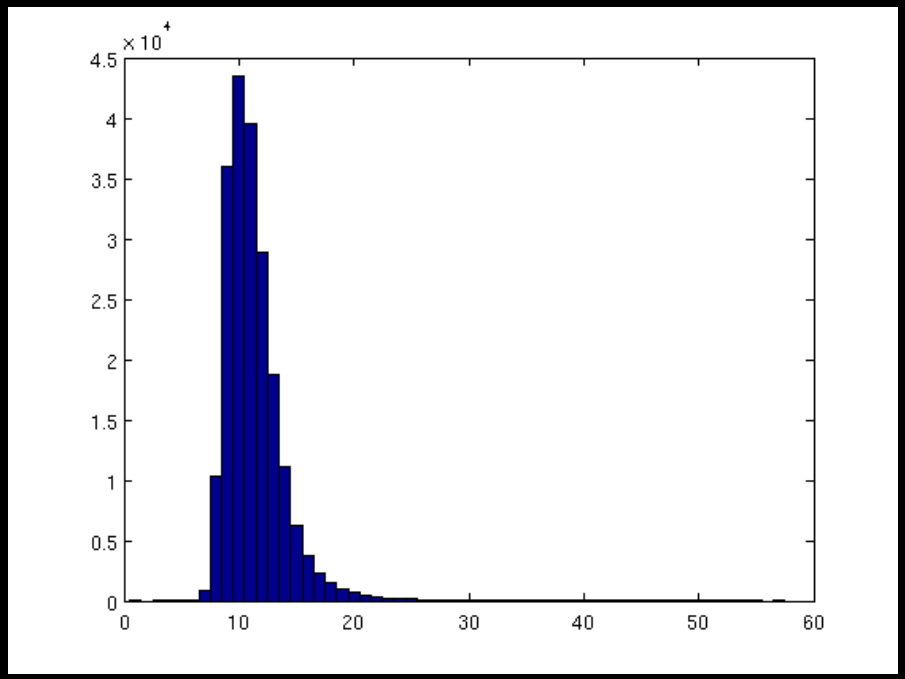


Real images: plug and play

- Datasets
  - PASCAL
  - Flickr 8K/30K
  - MS COCO
- Image Representation
  - Whole frame 4096D CNN features
- Text Preprocessing
  - Tokenization by Stanford CoreNLP
  - Lowercase all letters

Sentence Generation

- Procedure
  - Sample sentence length, n
  - Using language model, sample 100 sentences of length n
  - Select sentence with best likelihood



Length Distribution on COCO

Qualitative Results

PC: A plate with a bowl of soup and a cup of coffee on a table .

GT: A cup of coffee and a sandwich on a plate .

Qualitative Results

PC: An old brick building with a clock on top of it.

GT: A black and tan clock on a brown brick building.

Qualitative Results

PC: A cat sitting on top of a bed with a laptop.

GT: A dog on top of a person so you can not see the person.

Qualitative Results

PC: A bunch of stuffed animals sitting next to each other.

GT: An overhead view of a city cross walk on a rainy day features an array of colorful umbrellas.

Qualitative Results

PC: A laptop computer sitting on top of a desk with a laptop.

GT: An open laptop computer sitting on top of a white table.

Quantitative

	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

Conclusions

- CNN makes things work
- Language Models are nice to explore for language generation
- Try to reconstruct the image is helpful

Future Work

- Capture more details (better feature localization)
- Replace RNN with LSTM

THANKS! QUESTIONS?