Multimodal Learning Using Recurrent Neural Networks

Junhua Mao
mjhustc@ucla.edu
UCLA
10/18/2016

This talk follows from joint work with Alan Yuille, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Haoyuan Gao, Jonathan Huang, Kevin Murphy, Jiajing Xu, and Kevin Jing, among others.

Content

- The m-RNN image captioning model
  - Image caption generation
  - Image retrieval (given query sentence)
  - Sentence retrieval (given query image)
- Extensions
  - Incremental novel concept captioning
  - Multimodal word embedding learning
  - Referring expressions
  - Visual Question Answering

The m-RNN Image Captioning Model

http://www.stat.ucla.edu/~junhua.mao/m-RNN.html

a close up of a bowl of food on a table

a train is traveling down the tracks in a city

a pizza sitting on top of a table next to a box of pizza

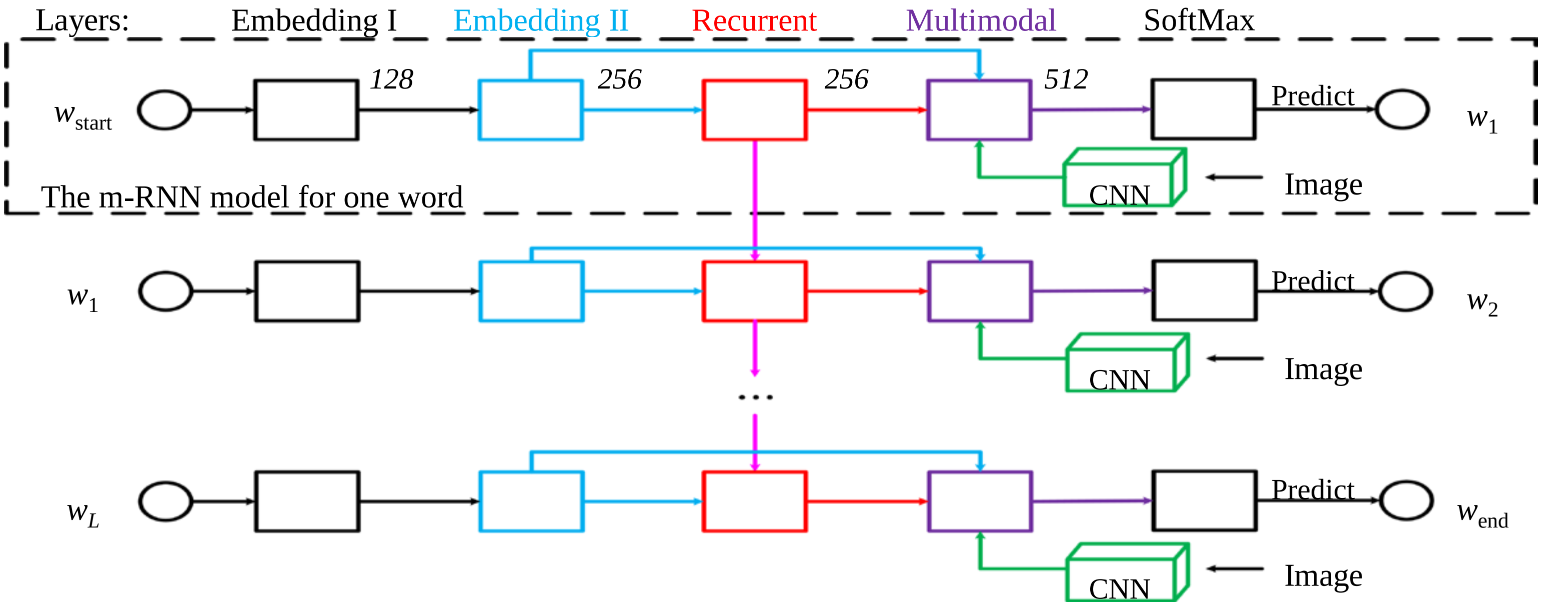
a cat laying on a bed with a stuffed animal

Mao, J., Xu, W., Yang, Y., Wang, J., Z. Huang & Yuille, A. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proc. ICLR 2015*.

Abstract

- Three Tasks:
  - Image caption generation
  - Image retrieval (given query sentence)
  - Sentence retrieval (given query image)
- One model (m-RNN):
  - A deep Recurrent NN (RNN) for the sentences
  - A deep Convolutional NN (CNN) for the images
  - A multimodal layer connects the first two components
- State-of-the-art Performance:
  - For three tasks
  - On four datasets: IAPR TC-12 [Grubinger et al. 06'], Flickr 8K [Rashtchian et al. 10'], Flickr 30K [Young et al. 14'] and MS COCO [Lin et al. 14']

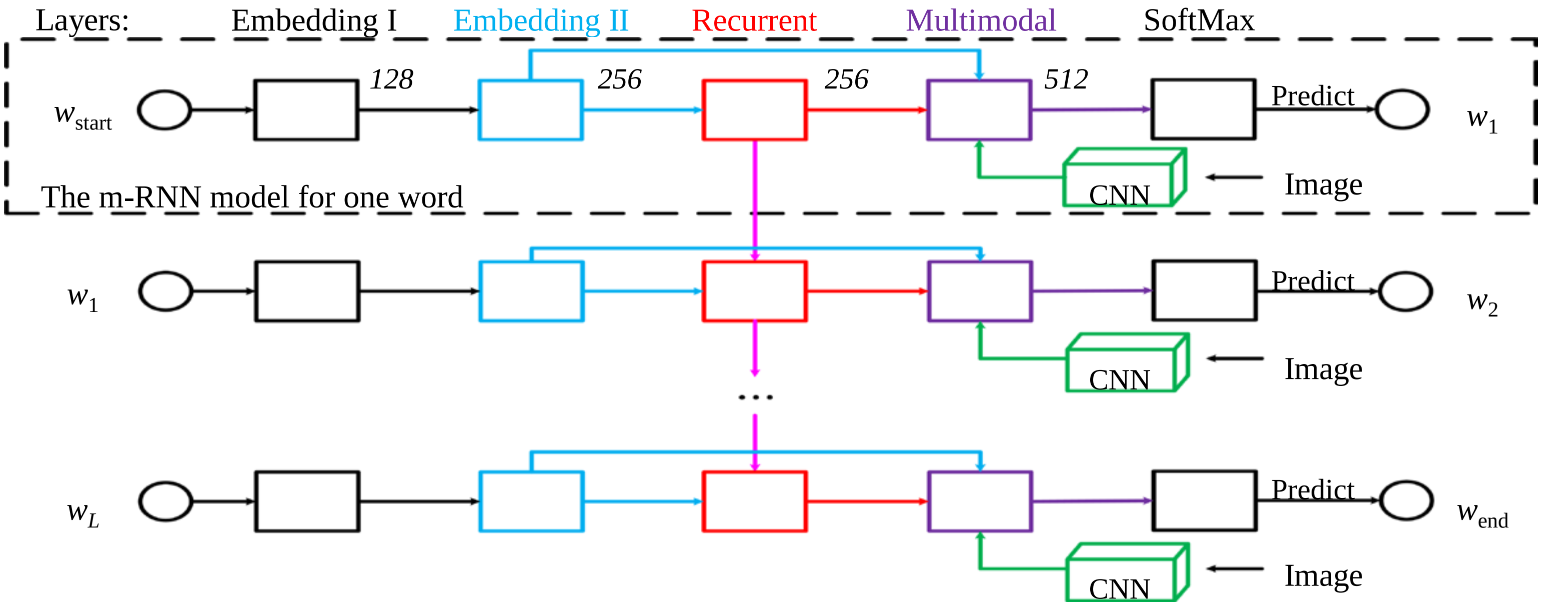
The m-RNN Model



$w_1, w_2, \ldots, w_L$ is the sentence description of the image

w_{start}, w_{end} is the start and end sign of the sentence

The m-RNN Model



The output of the trained model:

$$P(w_n | w_{1:n-1}, \mathbf{I})$$

Application

- Image caption generation:
  - Begin with the start sign w_{start}
  - Sample next word from $P(w_n|w_{1:n-1}, \mathbf{I})$
  - Repeat until generating w_{end}

Application

- Image caption generation:
  - Begin with the start sign w_{start}
  - Sample next word from $P(w_n|w_{1:n-1}, \mathbf{I})$
  - Repeat until generating w_{end}
- Image retrieval given query sentence:
  - Ranking score: $P\left(w_{1:L}^{Q}\middle|\mathbf{I}^{D}\right) = \prod_{n=2}^{L} P\left(w_n^Q\middle|w_{1:n-1}^{Q}, \mathbf{I}^{D}\right)$
  - Output the top ranked images

Application

- Image caption generation:
  - Begin with the start sign w_{start}
  - Sample next word from $P(w_n|w_{1:n-1}, \mathbf{I})$
  - Repeat until generating w_{end}

- Image retrieval given query sentence:
  - Ranking score: $P\left(w_{1:L}^{Q}\middle|\mathbf{I}^{D}\right) = \prod_{n=2}^{L} P\left(w_{n}^{Q}\middle|w_{1:n-1}^{Q}, \mathbf{I}^{D}\right)$
  - Output the top ranked images

- Sentence retrieval given query image:
  - *Challenge*: Some sentences have high probability for any image query
  - *Solution*: **Normalize** the probability. $\mathbf{I}'$ are sampled images:

$$\frac{P\left(w_{1:L}^{D}\middle|\mathbf{I}^{Q}\right)}{P\left(w_{1:L}^{D}\right)} \qquad P\left(w_{1:L}^{D}\right) = \sum_{\mathbf{I}'} P\left(w_{1:L}^{D}\middle|\mathbf{I}'\right) \cdot P(\mathbf{I}')$$

Application

- Image caption generation:
  - Begin with the start sign w_{start}
  - Sample next word from $P(w_n | w_{1:n-1}, \mathbf{I})$
  - Repeat until generating w_{end}

- Image retrieval given query sentence:
  - Ranking score: $P\left(w_{1:L}^{Q} \middle| \mathbf{I}^{D}\right) = \prod_{n=2}^{L} P\left(w_{n}^{Q} \middle| w_{1:n-1}^{Q}, \mathbf{I}^{D}\right)$
  - Output the top ranked images

- Sentence retrieval given query image:
  - *Challenge*: Some sentences have high probability for any image query
  - *Solution*: **Normalize** the probability. $\mathbf{I}'$ are sampled images:

$$\frac{P\left(w_{1:L}^{D} \middle| \mathbf{I}^{Q}\right)}{P\left(w_{1:L}^{D}\right)} \qquad P\left(w_{1:L}^{D}\right) = \sum_{\mathbf{I}'} P\left(w_{1:L}^{D} \middle| \mathbf{I}'\right) \cdot P(\mathbf{I}') \quad \xRightarrow[\text{Equivalent}]{} \quad P\left(\mathbf{I}^{Q} \middle| w_{1:L}^{D}\right) = \frac{P\left(w_{1:L}^{D} \middle| \mathbf{I}^{Q}\right) \cdot P(\mathbf{I}^{Q})}{P\left(w_{1:L}^{D}\right)}$$

Experiment: Retrieval

R@K: The recall rate of the groundtruth among the top K retrieved candidates

Med r: Median rank of the top-ranked retrieved groundtruth

	Sentence Retrival (Image to Text)				Image Retrival (Text to Image)			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Flickr30K								
Random	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeepFE-RCNN (Karpathy et al. 14')	16.4	40.2	54.7	8	10.3	31.4	44.5	13
RVR (Chen & Zitnick 14')	12.1	27.8	47.8	11	12.7	33.1	44.9	12.5
MNLM-AlexNet (Kiros et al. 14')	14.8	39.2	50.9	10	11.8	34.0	46.3	13
MNLM-VggNet (Kiros et al. 14')	23.0	50.7	62.9	5	16.8	42.0	56.5	8
NIC (Vinyals et al. 14')	17.0	56.0	/	7	17.0	**57.0**	/	7
LRCN (Donahue et al. 14')	14.0	34.9	47.0	11	/	/	/	/
DeepVS-RCNN (Karpathy et al. 14')	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2
Ours-**m-RNN**-AlexNet	18.4	40.2	50.9	10	12.6	31.2	41.5	16
Ours-**m-RNN**-VggNet	**35.4**	**63.8**	**73.7**	**3**	**22.8**	50.7	**63.1**	**5**
MS COCO								
Random	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeepVS-RCNN (Karpathy et al. 14')	29.4	62.0	75.9	2.5	20.9	**52.8**	69.2	4
Ours-**m-RNN**-VggNet	**41.0**	**73.0**	**83.5**	**2**	**29.0**	42.2	**77.0**	**3**

(*) Results reported on 04/10/2015.

Experiment: Captioning

Results on the MS COCO test set

	B1	B2	B3	B4	CIDEr	$ROUGE_L$	METEOR
Human-c5 (**)	0.663	0.469	0.321	0.217	0.854	0.484	0.252
m-RNN-c5	0.668	0.488	0.342	0.239	0.729	0.489	0.221
m-RNN-beam-c5	0.680	0.506	0.369	0.272	0.791	0.499	0.225
Human-c40 (**)	0.880	0.744	0.603	0.471	0.910	0.626	0.335
m-RNN-c40	0.845	0.730	0.598	0.473	0.740	0.616	0.291
m-RNN-beam-c40	0.865	0.760	0.641	0.529	0.789	0.640	0.304

c5 and c40: evaluated using 5 and 40 reference sentences respectively.

"-beam" means that we generate a set of candidate sentences, and then selects the best one. (beam search)

(**) Provided in https://www.codalab.org/competitions/3221#results

(***) We evaluate it on the MS COCO evaluation server: https://www.codalab.org/competitions/3221

Discussion: Chinese captions

一个年轻的男孩坐在长椅上。

A young boy sitting on a bench.

一列火车在轨道上行驶。

A train running on the track.

一辆双层巴士停在一个城市街道上。

A double decker bus stop on a city street.

We acknowledge Haoyuan Gao and Zhiheng Huang from Baidu Research for designing the Chinese image captioning system

Novel Visual Concept Learning

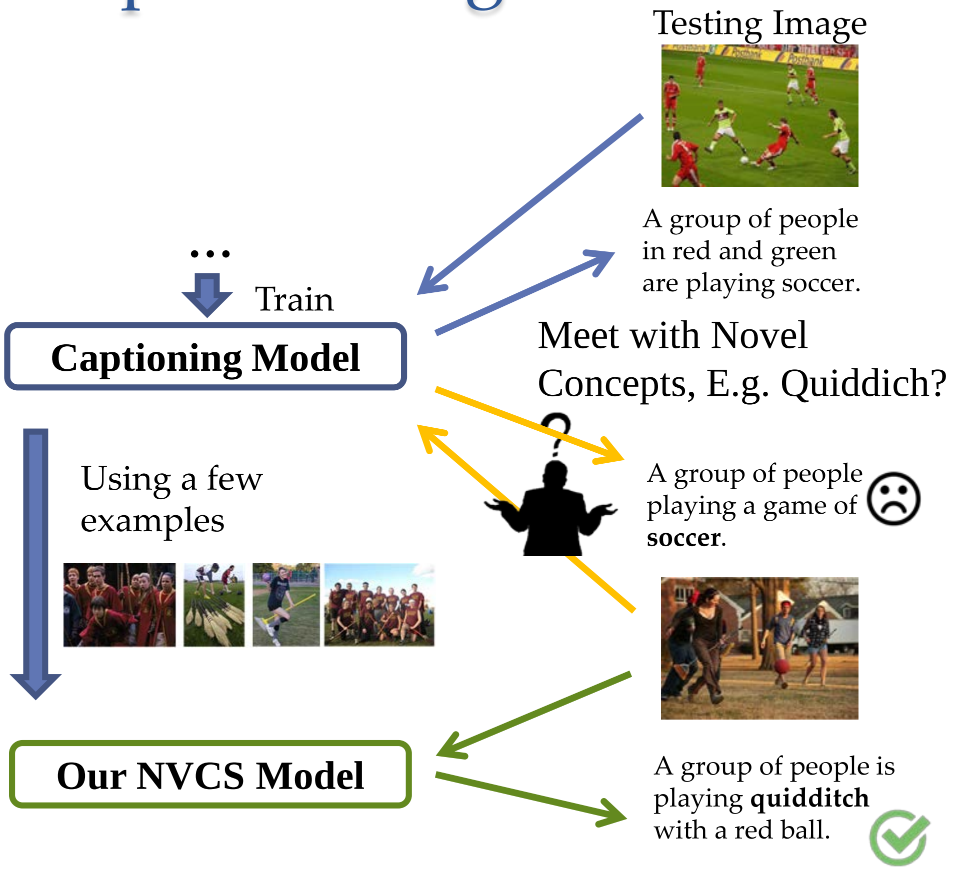
Novel Visual Concept Learning

The learning Novel Visual Concept from Sentences (*NVCS*) Task:

- Hard but important
  - o Slow if retrain the whole model
  - o Lack of training samples
  - o Overfit easily

Our contributions to this new task:

- Datasets [Released]
- A novel framework
  - o Fast, simple and effective
  - o No extensive retraining



Mao, J., Xu, W., Yang, Y., Wang, J., Z. Huang & Yuille, A. Learning like a Child: Fast Novel Visual Concept Learning from Sentences Descriptions. In *Proc. ICCV 2015*

The Novel Visual Concept Dataset

- Released on project page: http://www.stat.ucla.edu/~junhua.mao/projects/child_learning.html
- Contains 11 novel visual concepts
  - 100 images per concept, 5 sentences per image

Windmill / Quidditch

[illegible] flowers [illegible].

Rocket Gun / T-rex / Tempura

[illegible] plate.

Huangmei Opera / Sanshe / Water Cooler

[illegible] of trees

Wedding Dress / Kiss

[illegible] hands.

Our NVCS model & Results

Key components:

- Transposed Weight Sharing
  - Reduce ~50% parameters
- Baseline Probability Fixation
  - Avoid overfitting to novel concepts
  - Do not disturb previously learned concepts
  - Only need a few samples

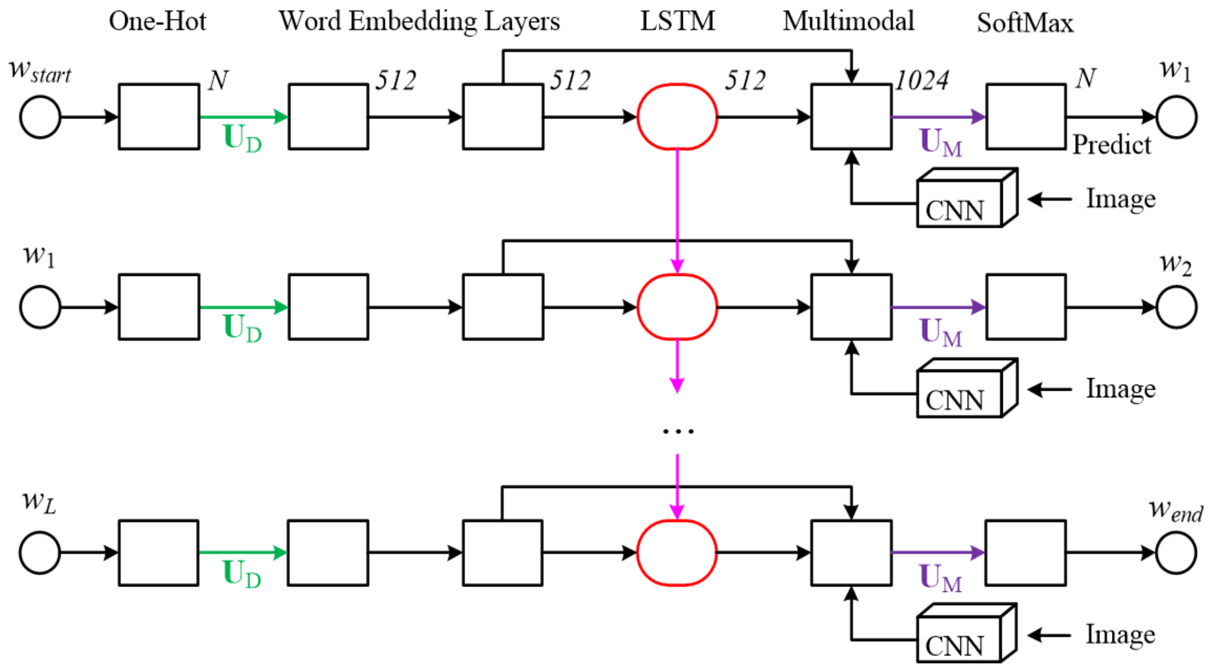
Compared to the re-training strategy from scratch

- Performs comparable or even better
- Much Faster (≥ 50 X speed up)

Example results:

Novel Concepts:	Cat	T-tex
		
Before NVCS Training	a window with a window	a man in blue shirt is riding a horse
After NVCS Training	a cat standing in front of a window	a t-rex is standing in a room with a man

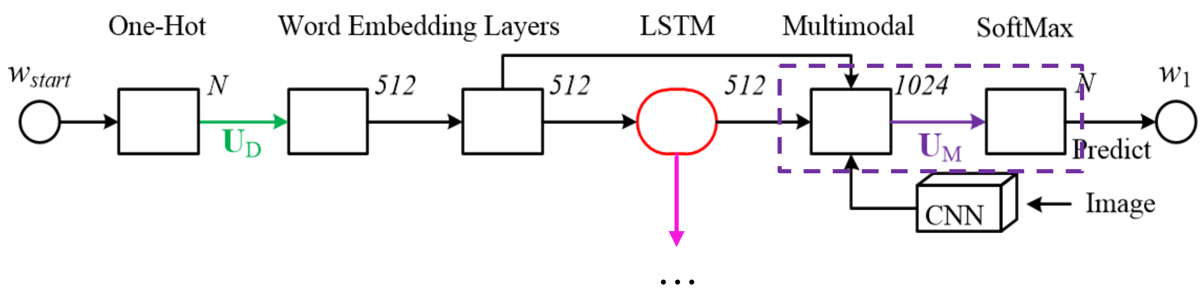
Transposed Weight Sharing



$$\mathbf{w}(t) = f(\mathbf{U}_\mathrm{D}\mathbf{h}(t))$$

$$\mathbf{y}(t) = g(\mathbf{U}_\mathrm{M}\mathbf{m}(t) + \mathbf{b})$$

Transposed Weight Sharing

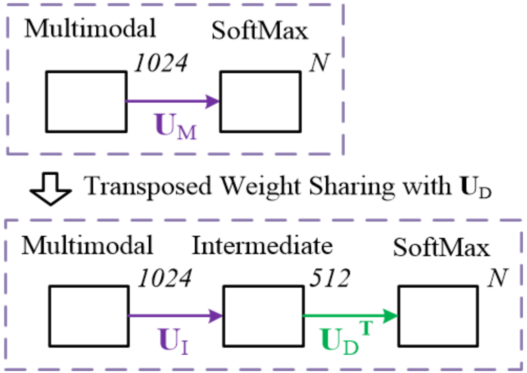


$$\mathbf{w}(t) = f(\mathbf{U}_\mathrm{D}\mathbf{h}(t))$$

$$\mathbf{y}(t) = g(\mathbf{U}_\mathrm{M}\mathbf{m}(t) + \mathbf{b})$$

Decompose $\mathrm{U_M}$ $\Downarrow$

$$\mathbf{y}(t) = g[\mathbf{U}_\mathrm{D}^T\ f(\mathbf{U}_\mathrm{I}\mathbf{m}(t)) + \mathbf{b}]$$



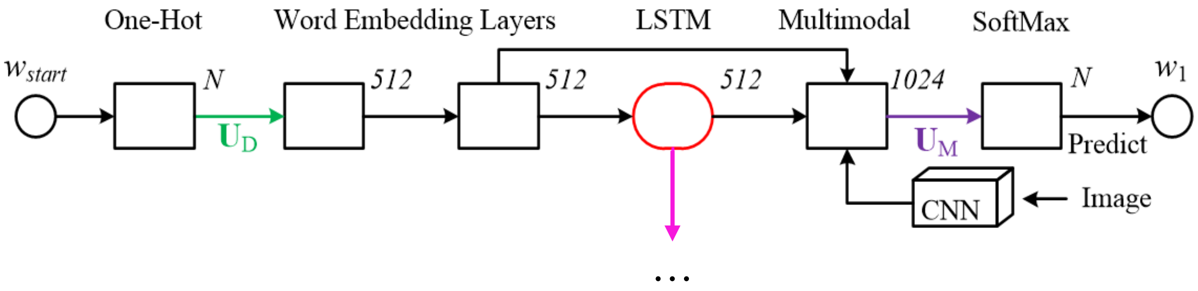
Transposed Weight Sharing

	B-1	B-2	B-3	B-4	METEOR	CIDEr	ROUGE_L
m-RNN [38]	0.680	0.506	0.369	0.272	0.225	0.791	0.499
ours-TWS	**0.685**	**0.512**	**0.376**	**0.279**	**0.229**	**0.819**	**0.504**

	BiasFix	Centralize	TWS	f
Deep-NVCS-UnfixedBias	$\times$	$\times$	$\surd$	0.851
Deep-NVCS-FixedBias	$\surd$	$\times$	$\surd$	0.860
Deep-NVCS-NoBPF-NoTWS	$\times$	$\times$	$\times$	0.839
Deep-NVCS-BPF-NoTWS	$\surd$	$\surd$	$\times$	0.850
Deep-NVCS-BPF-TWS	$\surd$	$\surd$	$\surd$	**0.875**

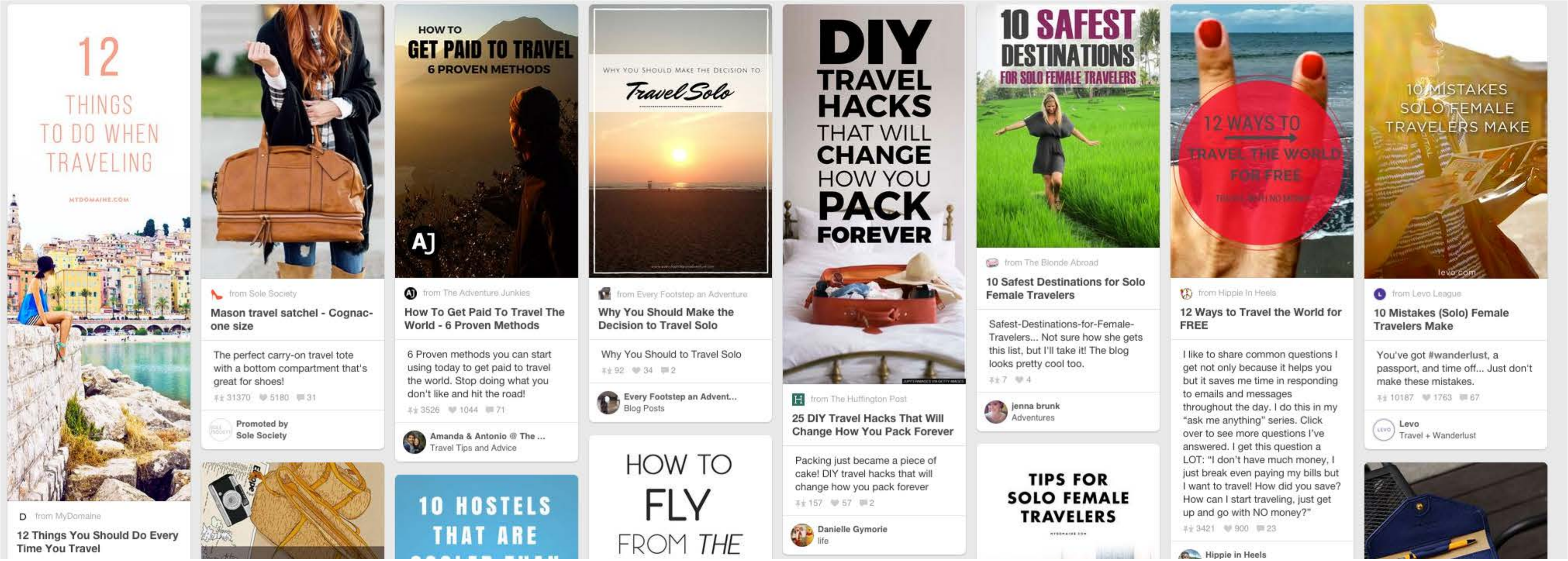
Discussion: Word Embeddings

New Word	Five nearest neighbours
cat	kitten; tabby; puppy; calico; doll;
motorcycle	motorbike; moped; vehicle; motor; motorbikes;
quidditch	soccer; football; softball; basketball; frisbees;
t-rex	giraffe's; bull; pony; goat; burger;
samisen	guitar; wii; toothbrushes; purse; contents;



Will learned word embedding benefit from the transposed weight sharing strategy?

Leaning Multimodal Word Embeddings



Pinterest

Mao, J., Xu, J., Jing, Y., & Yuille, A. Training and Evaluating Multimodal Word Embeddings with Large-scale Web Annotated Images. In *Proc. NIPS 2016*.

The Pinterest40M training dataset

This strawberry limeade cake is fruity, refreshing, and gorgeous! Those lovely layers are impossible to resist.

This is the place I will be going (hopefully) on my first date with Prince Stephen. It's the palace gardens, and they are gorgeous. I cannot wait to get to know him and exchange photography ideas!

White and gold ornate library with decorated ceiling, iron-work balcony, crystal chandelier, and glass-covered shelves. (I don't know if you're allowed to read a beat-up paperback in this room.)

This flopsy-wopsy who just wants a break from his walk. | 18 German Shepherd Puppies Who Need To Be Snuggled Immediately

Make two small fishtail braids on each side, then put them together with a ponytail.

The Pinterest40M training dataset

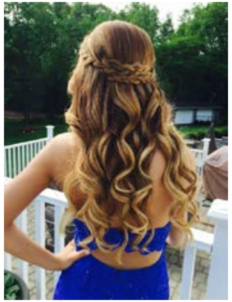
	Image	Sentences
Flickr8K [15]	8K	40K
Flickr30K [35]	30K	150K
IAPR-TC12 [12]	20K	34K
MS COCO [22]	200K	1M
Im2Text [28]	1M	1M
Pinterset40M	40M	300M

The Pinterest related phrase testing dataset

User Query

hair styles

Most Clicked Items



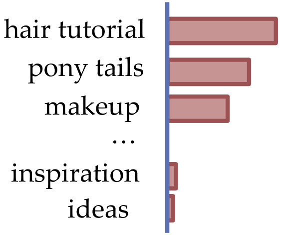
Annotations:
hair tutorial
prom night ideas
long hair
beautiful
…

Annotations:
pony tails
unique hairstyles
hair tutorial
picture inspiration
…

…

Positive Phrases for the Query

Sorted by Click



Remove overlapping words

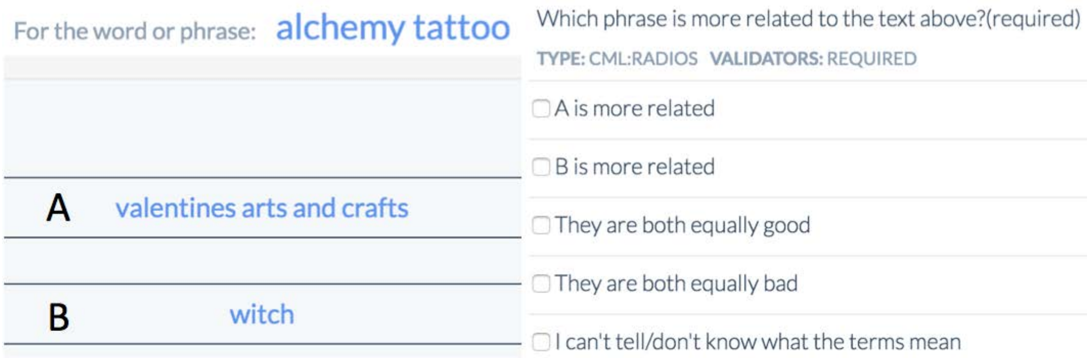
Final List

pony tails
makeup
…

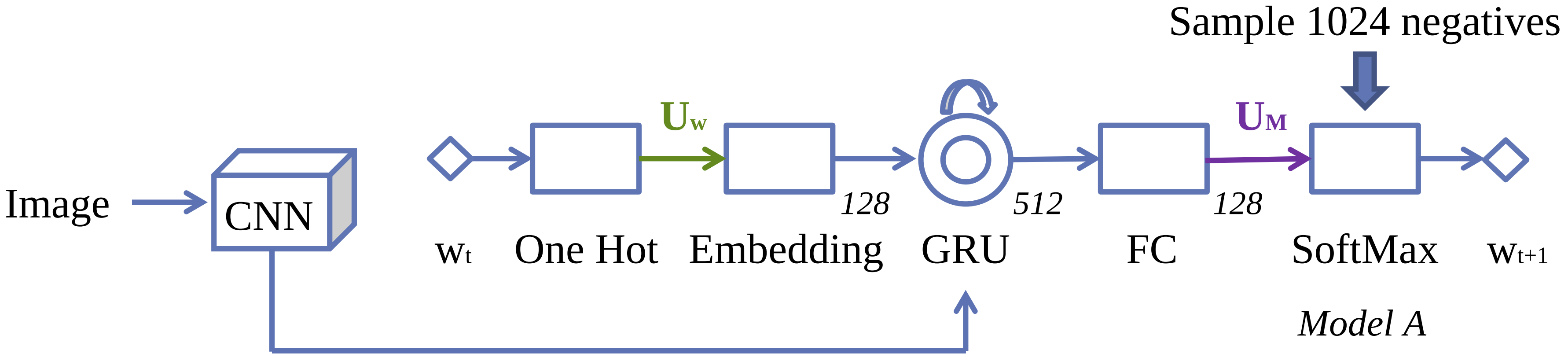
Related Phrase 10M dataset (RP10M)

Crowdsourcing cleaned up ⇩

Gold Related Phrase 10K dataset (Gold RP10K)



The Multimodal Word Embedding Models



Baselines (No transposed weight sharing):

Model B: Supervision on the final GRU state $$\mathcal{L}_{state} = \frac{1}{n}\sum_s \| h_{l_s} - \mathrm{ReLU}(W_I f_{I_s}) \|$$

Model C: Supervision on the embeddings $$\mathcal{L}_{emb} = \frac{1}{n}\sum_s \frac{1}{l_s}\sum_t \| e_t - \mathrm{ReLU}(W_I f_{I_s}) \|$$

Experiments

	Gold RP10K	RP10M	dim
Model A without visual (Pure text RNN)	0.748	0.633	128
Model A without weight sharing	0.773	0.681	128
Model A (weigh shared multimodal RNN)	**0.843**	**0.725**	128
Model B (direct visual supervisions on the final RNN state)	0.705	0.646	128
Model C (direct visual supervisions on the embeddings)	0.771	0.687	128
Word2Vec-GoogleNews [25]	0.716	0.596	300

Experiments

Content

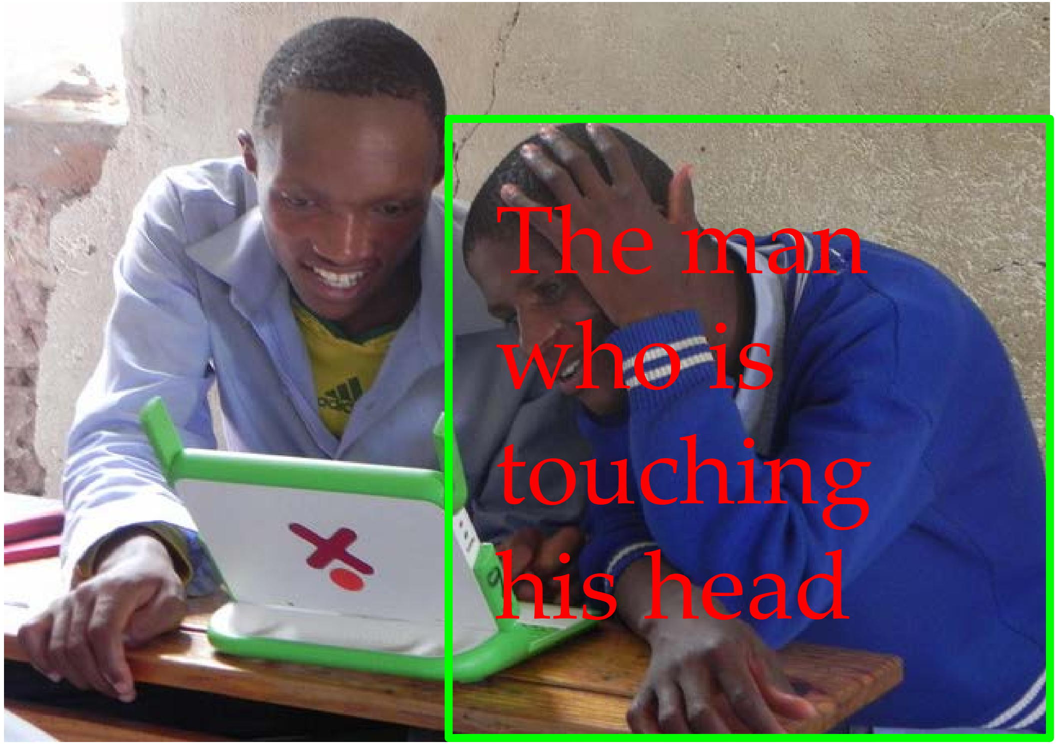
- The m-RNN image captioning model
  - Image caption generation
  - Image retrieval (given query sentence)
  - Sentence retrieval (given query image)
- Extensions
  - Incremental novel concept captioning
  - Multimodal word embedding learning
  - Referring expressions
  - Visual Question Answering

Unambiguous Object Descriptions

- ✗ A man.
- ✗ A man in blue.
- ✓ A man in blue sweater.
- ✓ A man who is touching his head.

Unambiguous Object Descriptions (Referring Expressions [Kazemzadeh *et.al* 2014]):

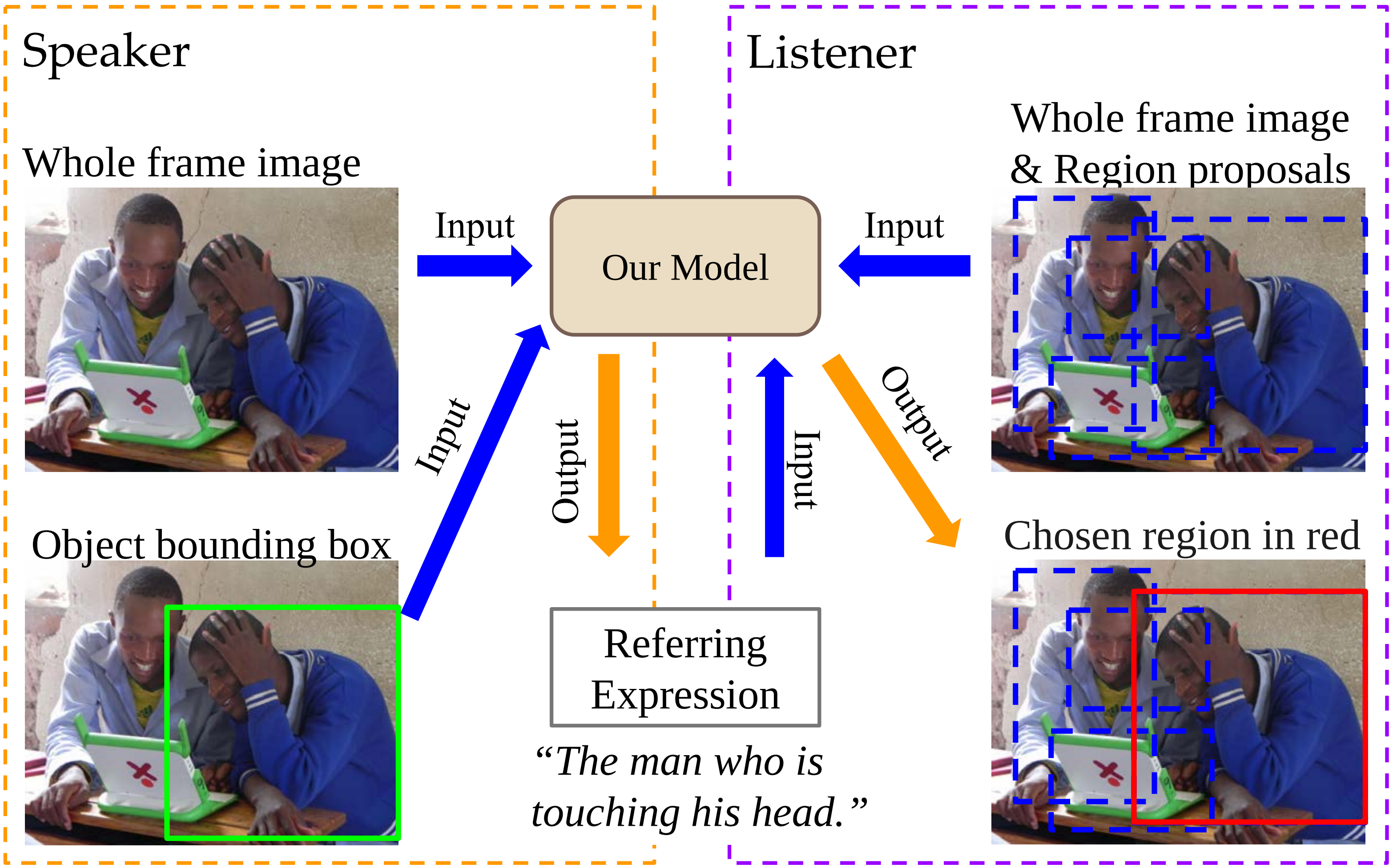
Uniquely describes the relevant object or region within its context.



It is hard to evaluate image captions.

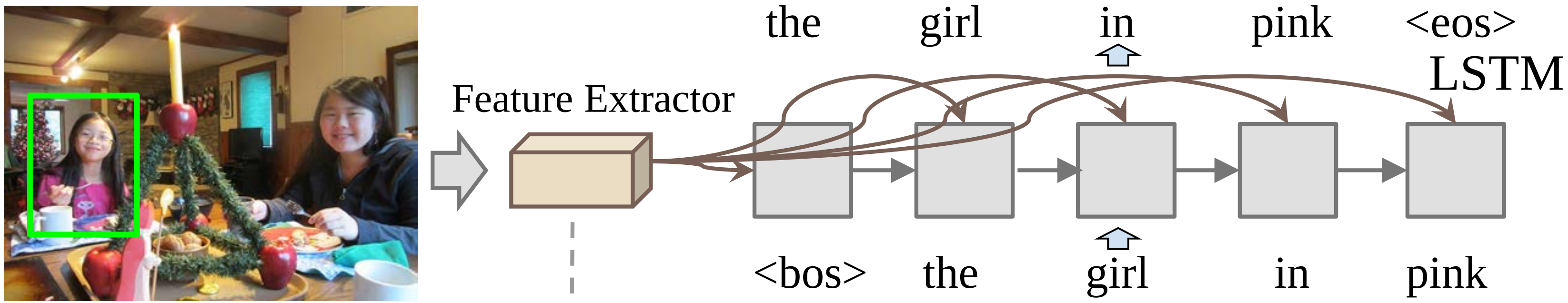
Two men are sitting next to each other.

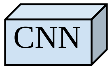
Two men are sitting next to each other in front of a desk watching something from a laptop.


Speaker
Whole frame image
Input
Our Model
Input
Listener
Whole frame image
& Region proposals
Input
Output
Input
Output
Object bounding box
Chosen region in red
Referring
Expression
"The man who is
touching his head."

The Baseline Model

Adapting a LSTM image captioning model ([Mao *et.al* 2015] [Vinyals *et. al* 2015] [Donahue *et.al* 2015] …)

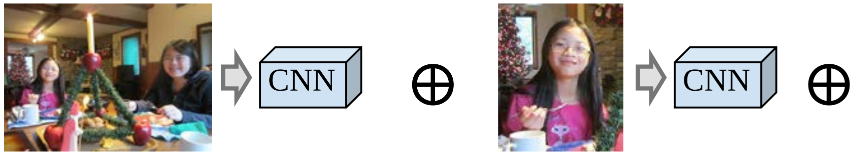




VGGNet
[Simonyan *et. al 2015*]

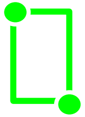
Training Objective: $J(\theta) = -\sum_{n=1}^{N} \log p(S_n | R_n, I_n, \theta)$

Feature Extractor (2005 dimension feature)



$[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{S_{bbox}}{S_{image}}]$

(x_{tl}, y_{tl})



(x_{br}, y_{br})

The Speaker-Listener Pipeline

Speaker module:

1. Decode with beam search
2. Hard to evaluate by itself?

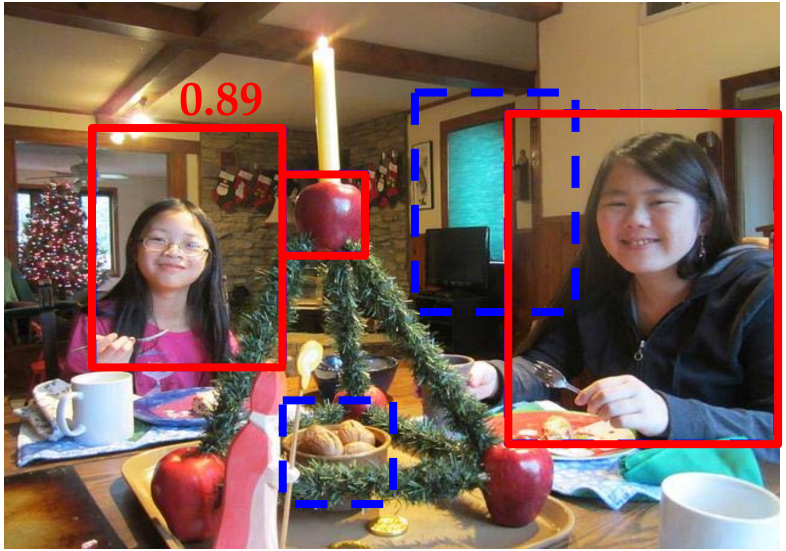
The Speaker-Listener Pipeline

Speaker module:

1. Decode with beam search
2. Hard to evaluate by itself?

Listener module:

"A girl in pink"



Multibox Proposals [Erhan *et.al* 2014]

$p(S|R_n, I)$

0.89

0.32

0.05

...

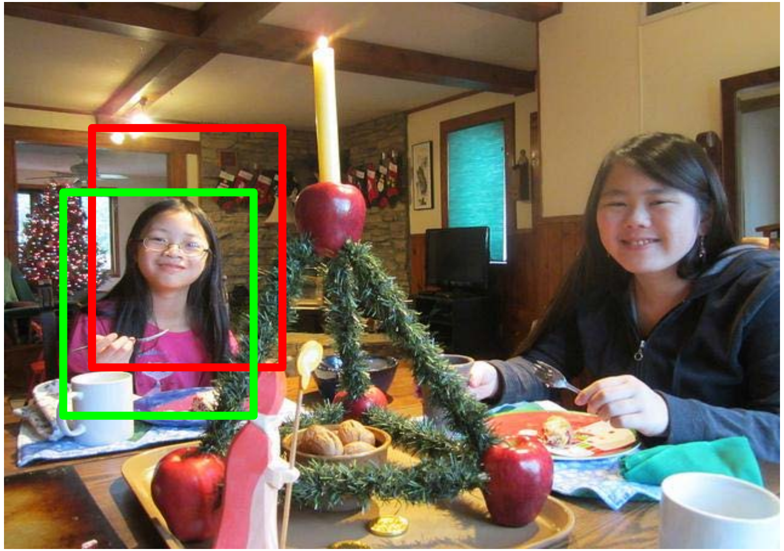
The Speaker-Listener Pipeline

Speaker module:

1. Decode with beam search
2. Hard to evaluate by itself?

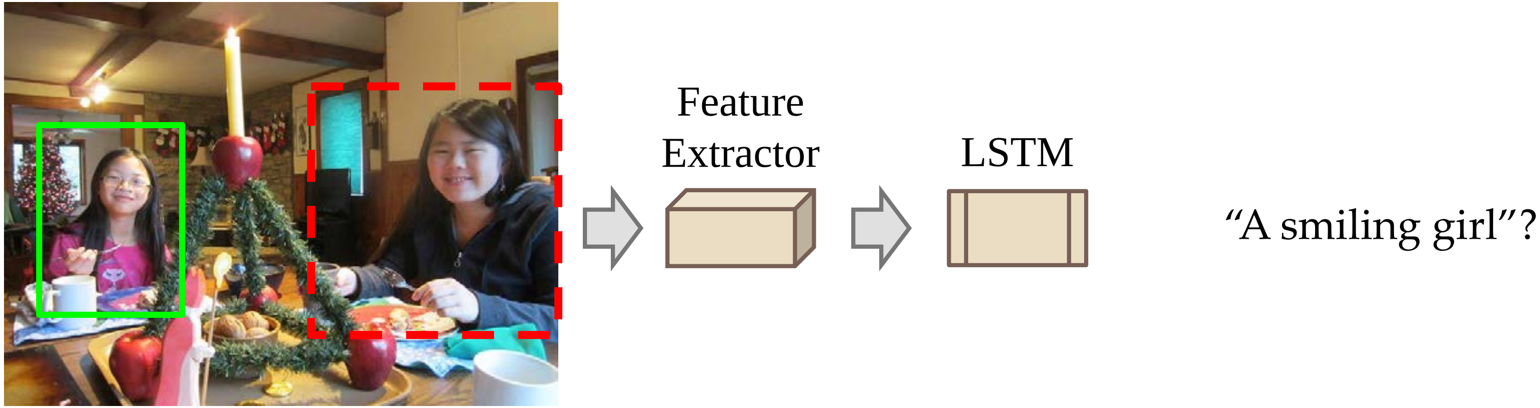
Listener module:

"A girl in pink"



- Easy to objectively evaluate.
  - *Precision@1*
- Evaluate the whole system in an end-to-end way

Speaker needs to *consider the listener*



The baseline model want to maximize the $p(S|R, I)$

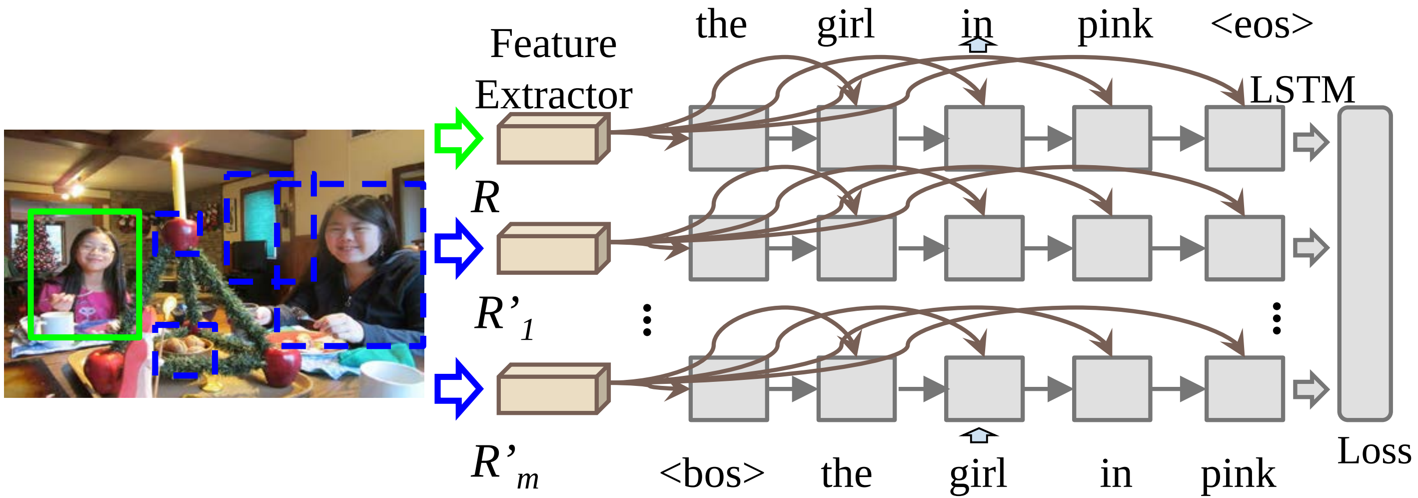
- Good to generate generic descriptions
- Not discriminative

A better, more discriminative model:

- Consider all possible regions
- maximize the gap between $p(S|R, I)$ and $p(S|R', I)$

R': regions serve as negatives for *R*

Our Full Model



Training Objective:

$$J'(\theta) = -\sum_{n=1}^{N} \log \frac{p(S_n|R_n, I_n, \theta)}{\sum_{R' \in C(I_n)} p(S_n|R', I_n, \theta)} = -\sum_{n=1}^{N} \log p(R_n|S_n, I_n, \theta)$$

$$J''(\theta) = -\sum_{n=1}^{N} \left\{ \log p(S_n|R_n, I_n, \theta) + \lambda \max\left(0, M - \log p(S_n|R_n, I_n, \theta) + \log p(S_n|R'_n, I_n, \theta)\right) \right\}$$

Dataset

Available at https://github.com/mjhucla/Google_Refexp_toolbox

The black and yellow backpack sitting on top of a suitcase.

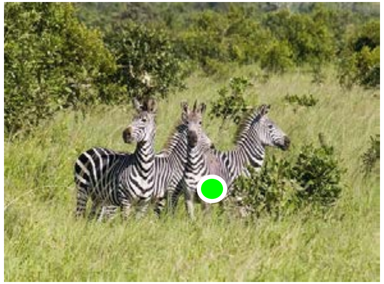
A yellow and black back pack sitting on top of a blue suitcase.

A boy brushing his hair while looking at his reflection.

A young male child in pajamas shaking around a hairbrush in the mirror.

An apple desktop computer.

The white IMac computer that is also turned on.



Zebra looking towards the camera.

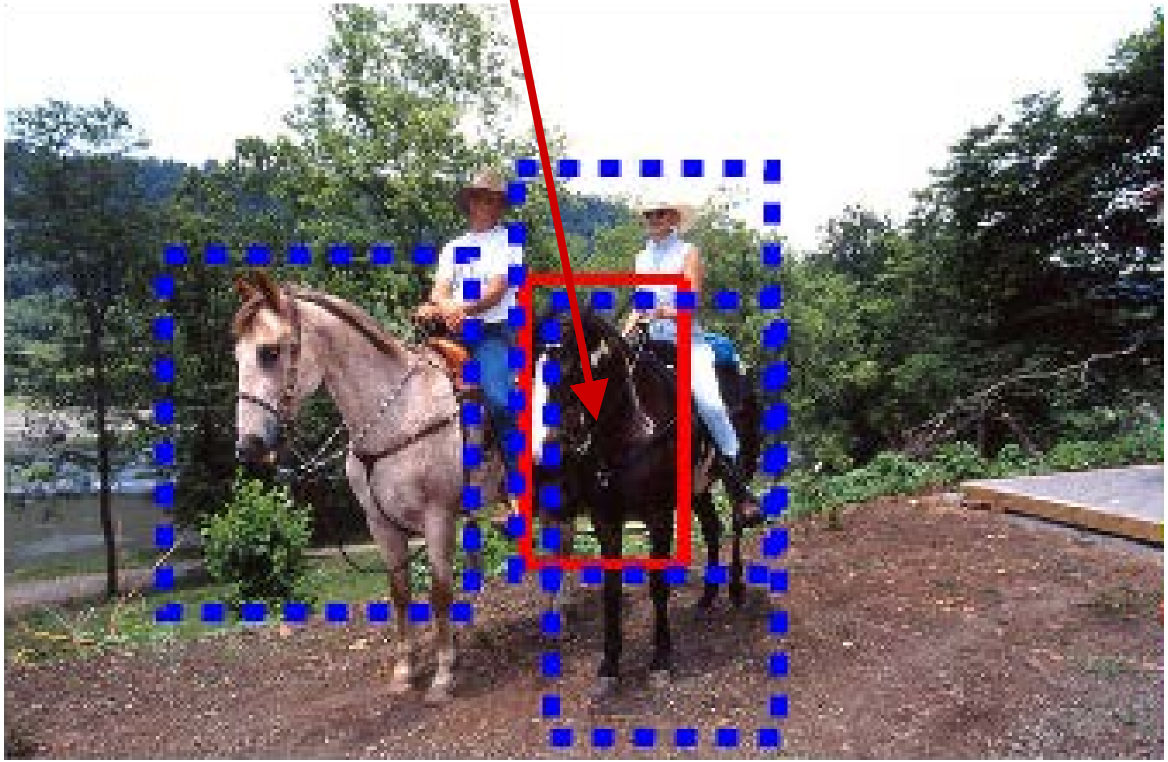
A zebra third from the left.

26,711 images (from MS COCO [Lin et.al 2015]), 54,822 objects and 104,560 referring expressions.

Listener Results

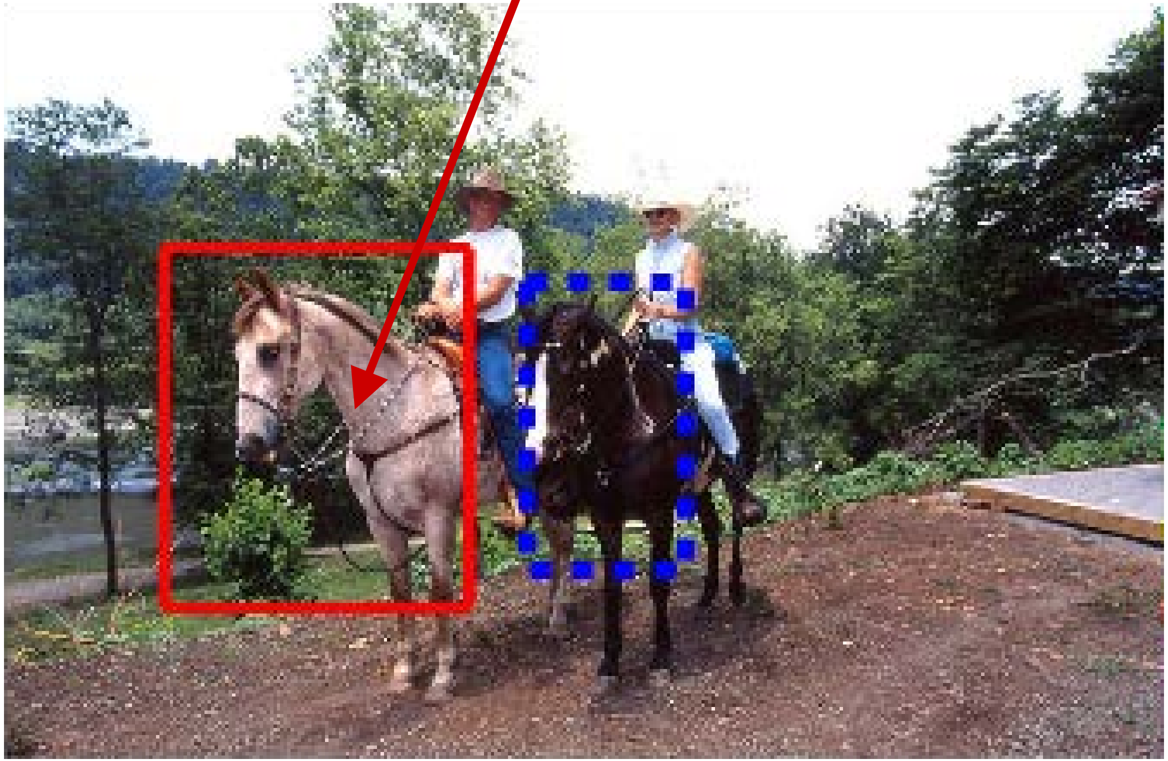
A dark brown horse with a white stripe wearing a black studded harness.

A dark brown horse with a white stripe wearing a black studded harness.



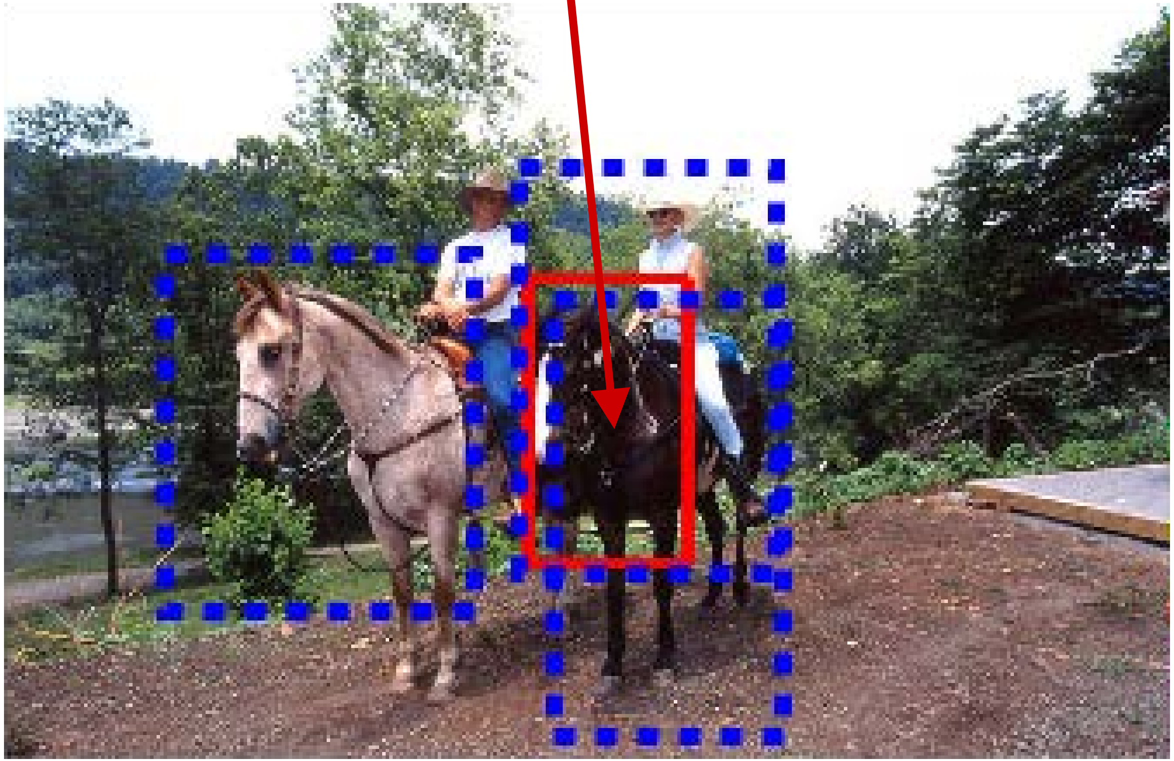
A white horse carrying a man.

A white horse carrying a man.



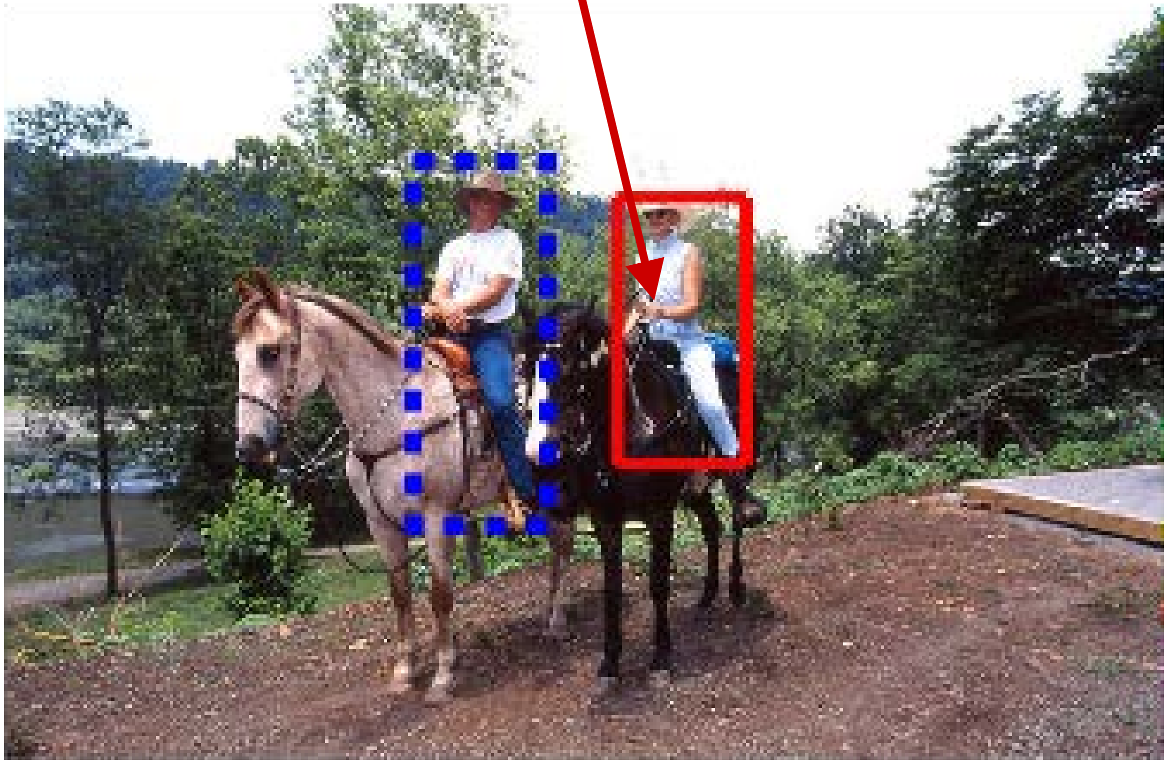
A dark horse carrying a woman

A dark horse carrying a woman

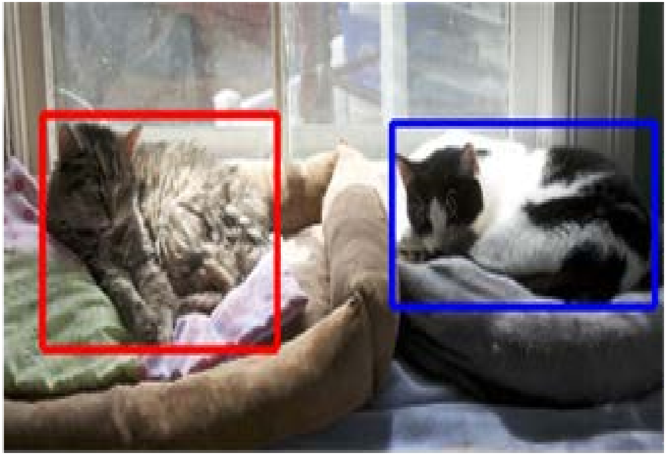


A woman on the dark horse.

A woman on the dark horse.



Speaker Results



A cat laying on a bed.

A black and white cat.

Baseline

A cat laying on the left.

A black cat laying on the right.

Our Full Model

Experiments: 4% improvement for precision@1

	Baseline	Our full model	Improvement
Listener Task	40.6	44.6	4.0
End-to-End (Speaker & Listener)	48.5	51.3	2.8
Human Evaluation (Speaker Task)	15.9	20.4	4.5

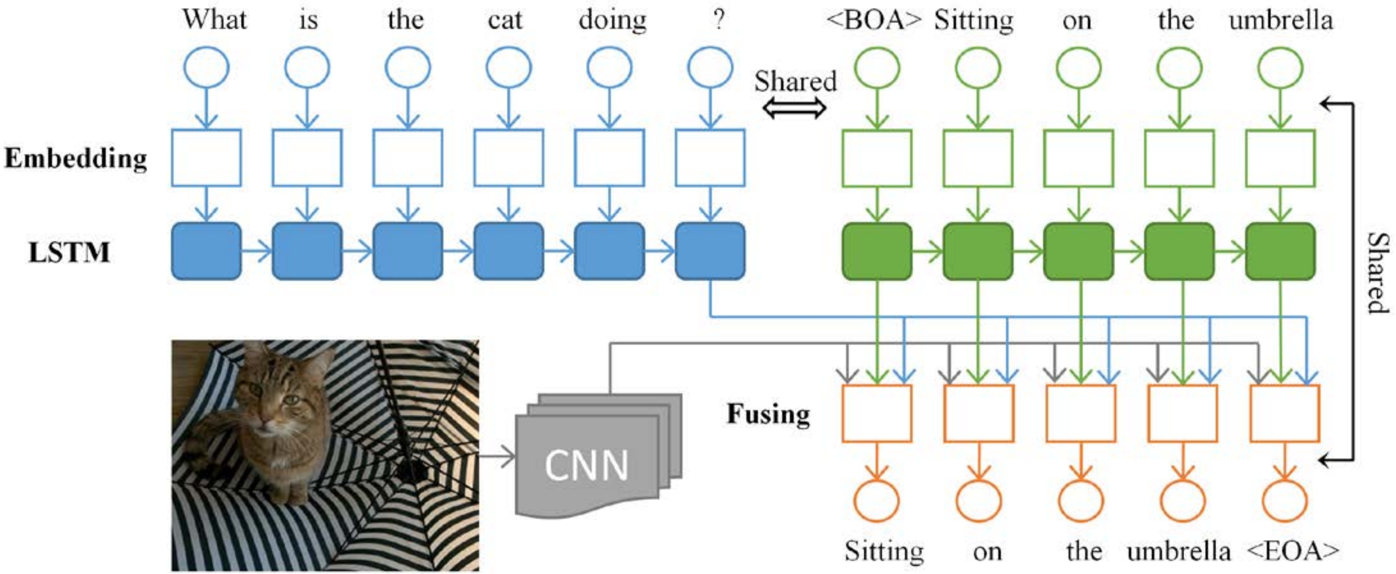
Take Home Message

To be a better communicator, need to be a better listener!!

Content

- The m-RNN image captioning model
  - Image caption generation
  - Image retrieval (given query sentence)
  - Sentence retrieval (given query image)
- Extensions
  - Incremental novel concept captioning
  - Multimodal word embedding learning
  - Referring expressions
  - **Visual Question Answering**

Visual Question Answering



Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In *Proc. NIPS 2015*.

Visual Question Answering

Image



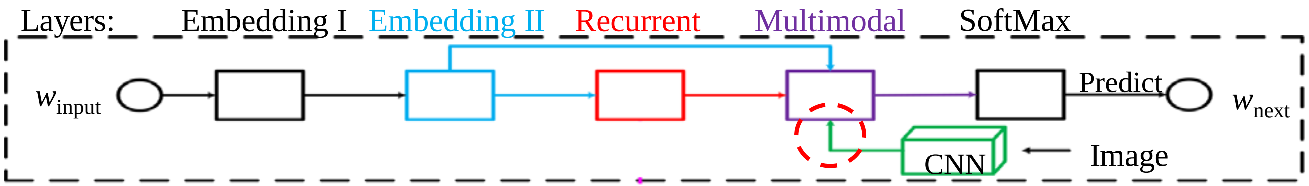
Question	公共汽车是什么颜色的？ What is the color of the bus?	黄色的是什么？ What is there in yellow?	草地上除了人以外还有什么动物？ What is there on the grass, except the person?	猫咪在哪里？ Where is the kitty?	观察一下说出食物里任意一种蔬菜的名字 ？ Please look carefully and tell me what is the name of the vegetables in the plate?
Answer	公共汽车是红色的。 The bus is red.	香蕉。 Bananas.	羊。 Sheep.	在椅子上 。 On the chair.	西兰花 。 Broccoli.

The results of the mQA model on FM-IQA dataset.

	Visual Turing Test			Human Rated Scores			
	Pass	Fail	Pass Rate (%)	2	1	0	Ave. Score
Human	948	52	94.8	927	64	9	1.918
blind-QA	340	660	34.0	-	-	-	-
mQA	647	353	64.7	628	198	174	1.454


THANK
YOU

Discussion: Nearest Image Search



$$\mathbf{m}(t) = g(\mathbf{V}_w \cdot \mathbf{w}(t) + \mathbf{V}_r \cdot \mathbf{r}(t) + \mathbf{V}_I \cdot \mathbf{I})$$

Refined Image Features