Multimodal Learning for Image Captioning and Visual Question Answering

Xiaodong He

Deep Learning Technology Center
Microsoft Research

UC Berkeley, April 7th, 2016

Collaborators:

Hao Fang
Saurabh Gupta
Forrest Iandola
Rupesh Srivastava
Li Deng
Piotr Dollár
Jianfeng Gao
Xiaodong He
Margaret Mitchell
John Platt
Lawrence Zitnick
Geoffrey Zweig
Jacob Devlin
Kenneth Tran
Lei Zhang
Jian Sun
Chris Buehler
Chris Thrasher
Chris Sienkiewicz
Cornelia Carapcea
Yuxiao Hu
Yandong Guo
Zichao Yang
Alex Smola
...

Outline

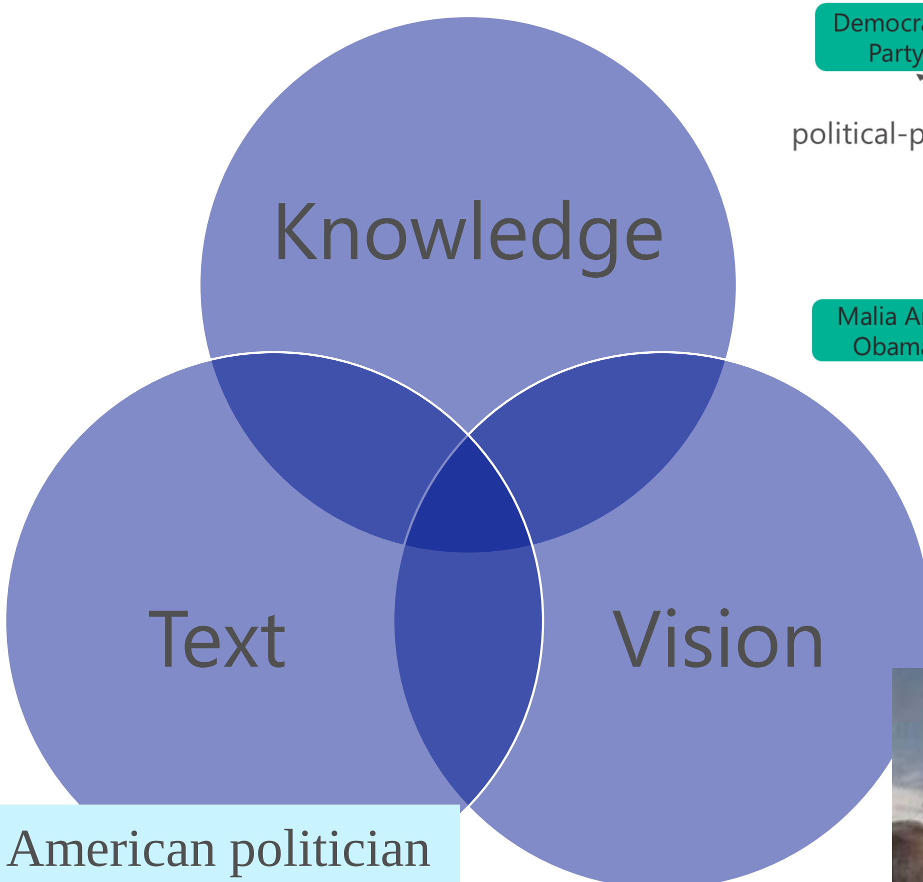
Motivation

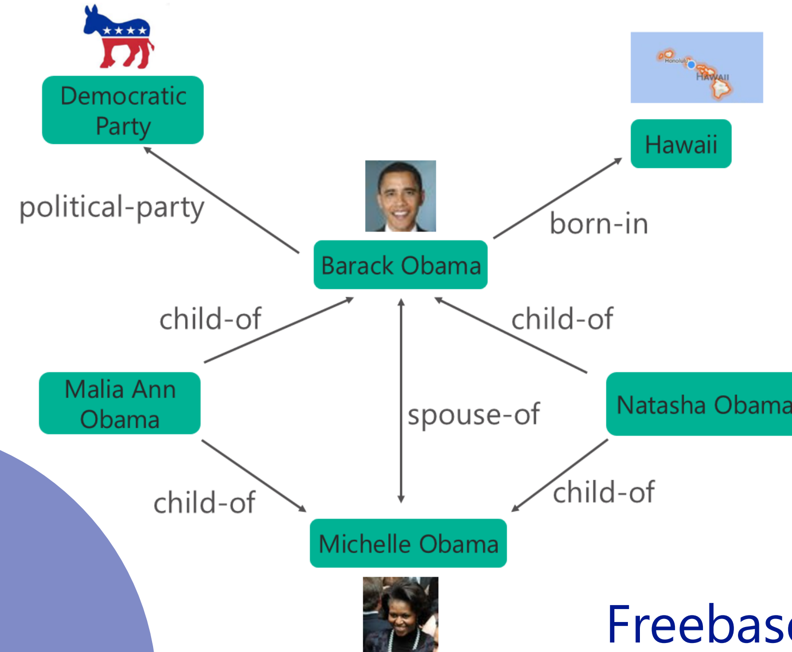
Image captioning

Visual question answering

Summary

Motivation: Humans learn to process text, image, and knowledge jointly





Freebase

Barack Obama is an American politician serving as the 44th President of the United States. Born in Honolulu, Hawaii, … in 2008, he defeated Republican nominee and was inaugurated as president on January 20, 2009.
(Wikipedia.org)

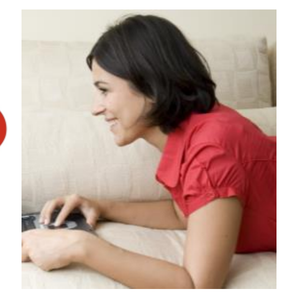
Image Captioning (one step from perception to cognition)

describe objects, attributes, and relationship in an image, in a natural language form

a man holding a tennis racquet on a tennis court

the man is on the tennis court playing a game

-- Let's do a Turing Test!

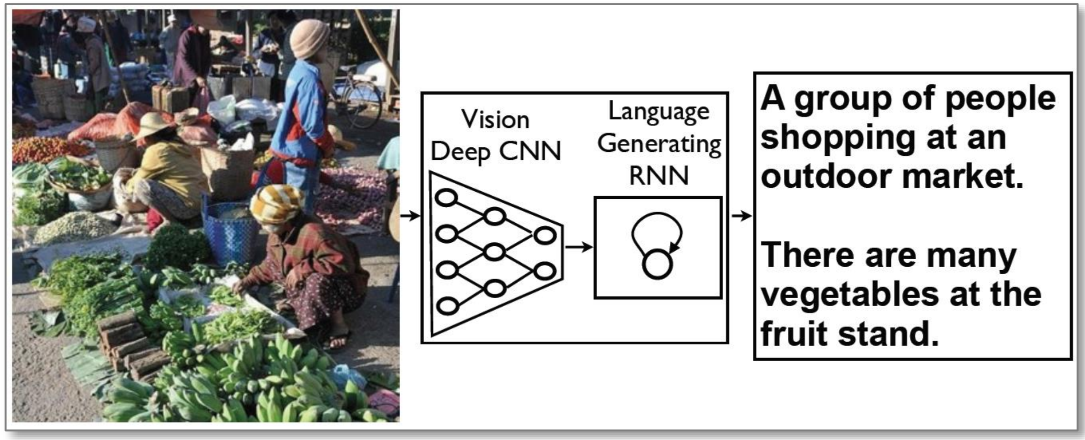


Two major paradigms

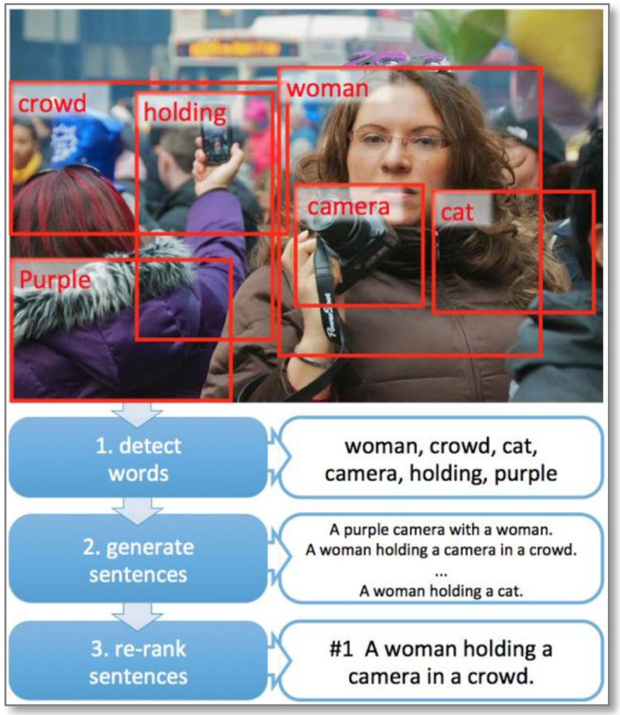
Two entries tied at the 1st place at COCO 2015 Caption Challenge

End-to-end using LSTM (e.g., Google)

Adopted **encoder-decoder** framework from machine translation, Popular: Google, Montreal, Stanford, Berkeley



Vinyals, Toshev, Bengio, Erhan, "**Show and Tell: A Neural Image Caption Generator**," CVPR, June 2015



Compositional framework (e.g., MSR)

Visual concept **detection** => caption **candidates generation** => Deep **semantic ranking**

Compositional framework can potentially exploit non paired image-caption data more effectively

[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "**From Captions to Visual Concepts and Back,**" CVPR, June 2015]

MSR, Stage 1: Multiple Instance Learning (MIL)

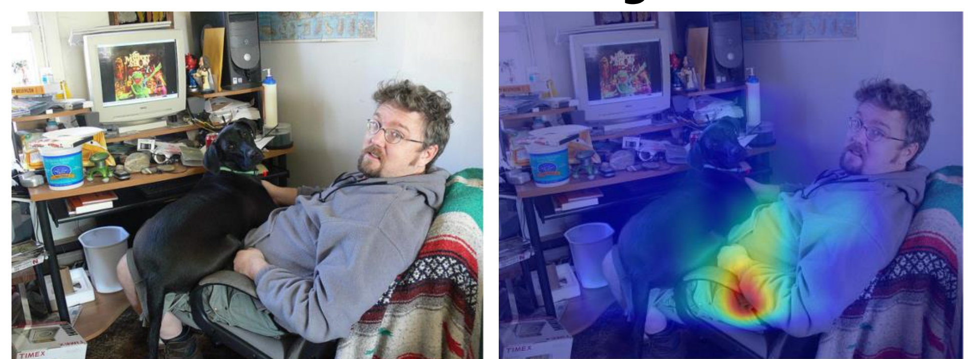
- Treat training caption as bag of image labels
- Train one binary classifier per label on all images
- "Noisy-Or" classifier
  - Image divided into 12x12 overlapping regions
  - fc7 vector used for image features

$$p_i^w = 1 - \prod_{j \in r_i} \left(1 - \overbrace{\sigma(f_{ij} \cdot v_w)}^{p(w\ in\ r_j\ of\ image\ i)}\right)$$

i= image id
f_{ij} = `fc7` vector
$\sigma(x)$ = sigmoid
r_i = regions
v_w = learned classifier weights

e.g., the visual "attention" of word **sitting**.

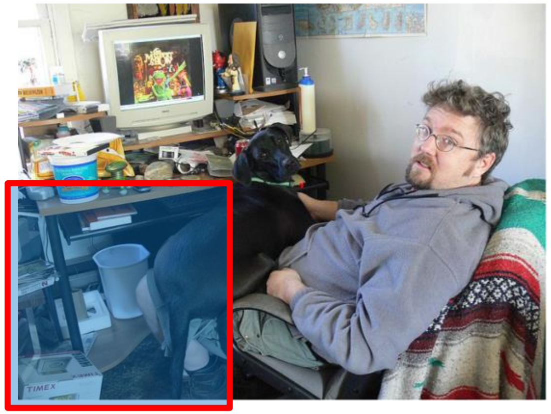
sitting



sitting (0.83)

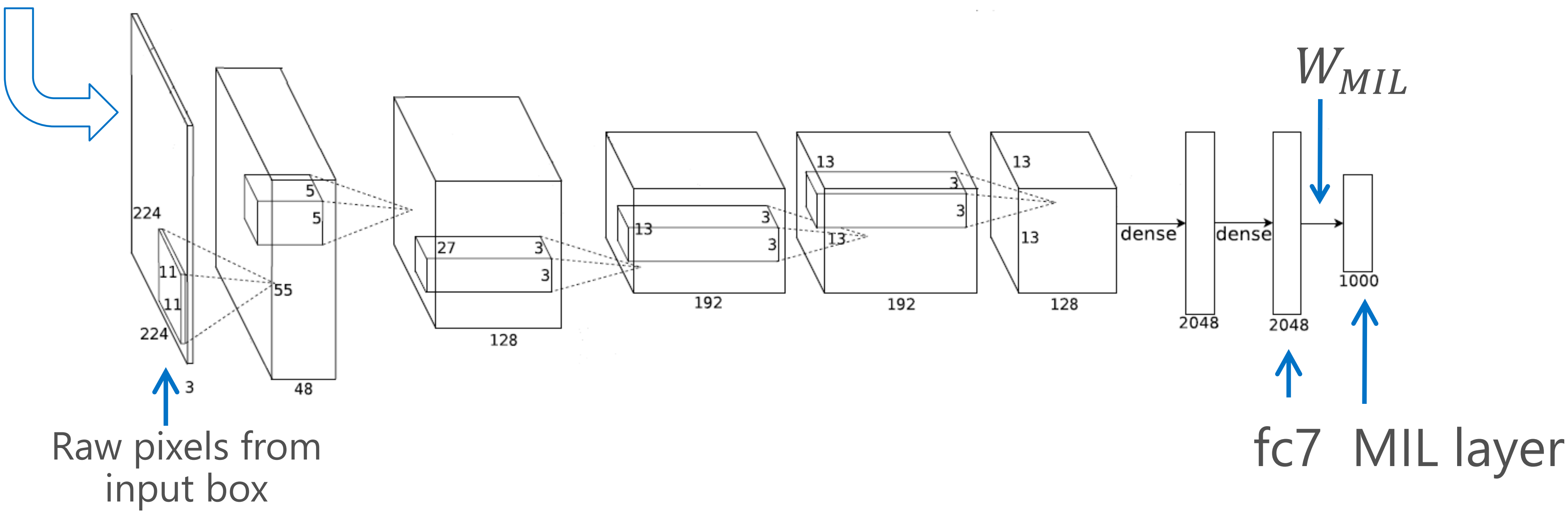
$$\mathrm{h}(x, y) = \sum_{r_i, s.t., (x,y) \in r_i} \sigma(f_{ij} \cdot v_{sitting})$$

Multiple Instance Learning illustration



a man sitting on a chair with a dog in his lap

$$\vec{P}(w\ in\ region) = 1/(1 + e^{W_{MIL} \times v_{fc7}})$$



Tuned image features from AlexNet (Krizhevsky et al., 2012) or VGG (Simonyan and Zisserman, 2014).

MaxEnt LM (MELM) for modeling language

Table 1. Features used in the maximum entropy language model.

Feature	Type	Definition	Description
Attribute	0/1	$\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	Predicted word is in the attribute set, i.e. has been visually detected and not yet used.
N-gram+	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is in the attribute set.
N-gram-	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \notin \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is not in the attribute set.
End	0/1	$\bar{w}_l = \kappa$ and $\tilde{\mathcal{V}}_{l-1} = \emptyset$	The predicted word is κ and all attributes have been mentioned.
Score	$\mathbb{R}$	score($\bar{w}_l$) when $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	The log-probability of the predicted word when it is in the attribute set.

$$\Pr(w_l = \bar{w}_l | \bar{w}_{l-1}, \cdots, \bar{w}_1, <s>, \tilde{\mathcal{V}}_{l-1}) = \frac{\exp\left[\sum_{k=1}^{K} \lambda_k f_k(\bar{w}_l, \bar{w}_{l-1}, \cdots, \bar{w}_1, <s>, \tilde{\mathcal{V}}_{l-1})\right]}{\sum_{v \in \mathcal{V} \cup </s>} \exp\left[\sum_{k=1}^{K} \lambda_k f_k(v, \bar{w}_{l-1}, \cdots, \bar{w}_1, <s>, \tilde{\mathcal{V}}_{l-1})\right]} \quad (3)$$

where $<s>$ denotes the start-of-sentence token, $\bar{w}_j \in \mathcal{V} \cup </s>$, and $f_k(w_l, \cdots, w_1, \tilde{\mathcal{V}}_{l-1})$ and λ_k respectively denote the k-th max-entropy feature and its weight. The basic discrete ME features we use are summarized in Table 1.

$$L(\Lambda) = \sum_{s=1}^{S} \sum_{l=1}^{\#(s)} \log \Pr(\bar{w}_l^{(s)} | \bar{w}_{l-1}^{(s)}, \cdots, \bar{w}_1^{(s)}, <s>, \tilde{\mathcal{V}}_{l-1}^{(s)}) \quad (4)$$

MELM for candidate generation

a kitchen with wooden

Language model

~~wooden~~ room ~~kitchen~~ sink stove ~~cabinets~~ floor

Image

cabinets

MaxEnt LM

$p(cabinets|with\ wooden)$

a kitchen with wooden cabinets

Repeat to generate 500 candidates

1. wooden cabinets in a kitchen
2. a sink and cabinets

...

500. a room with stove on the floor

[Fang, et al., CVPR 2015]

Deep Multimodal Similarity Model

- Project sentence and image into a comparable semantic vector space
- Whole sentence language model
- DMSM + basic features → re-ranked caption list

Q = image, D = caption, R = relevance

Relevance: $R(Q,D) = \text{cosine}(y_Q, y_D) = \frac{{y_Q}^T y_D}{\|y_Q\|\|y_D\|}$

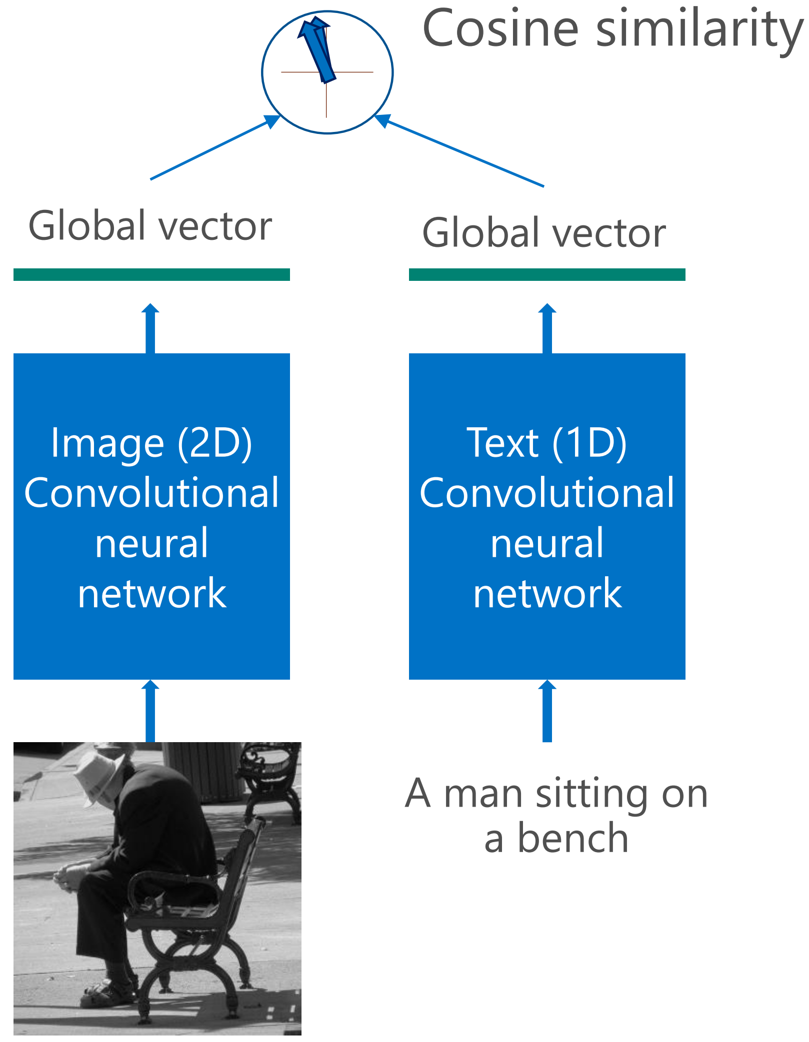
Caption probability: $P(D|Q) = \frac{\exp(\gamma R(Q,D))}{\Sigma_{D' \in \mathbf{D}} \exp(\gamma R(Q,D'))}$

Candidate captions

Smoothing factor

Objective: $L(\Lambda) = -\log \prod_{(Q,D^+)} P(D^+|Q)$

Correct caption

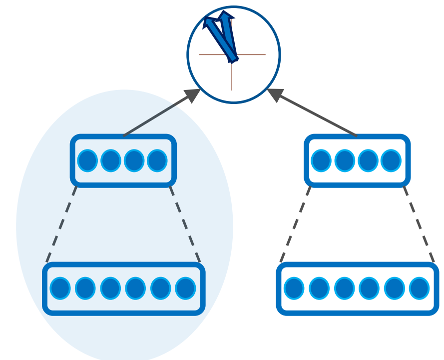


Serves as a semantic matching checker.

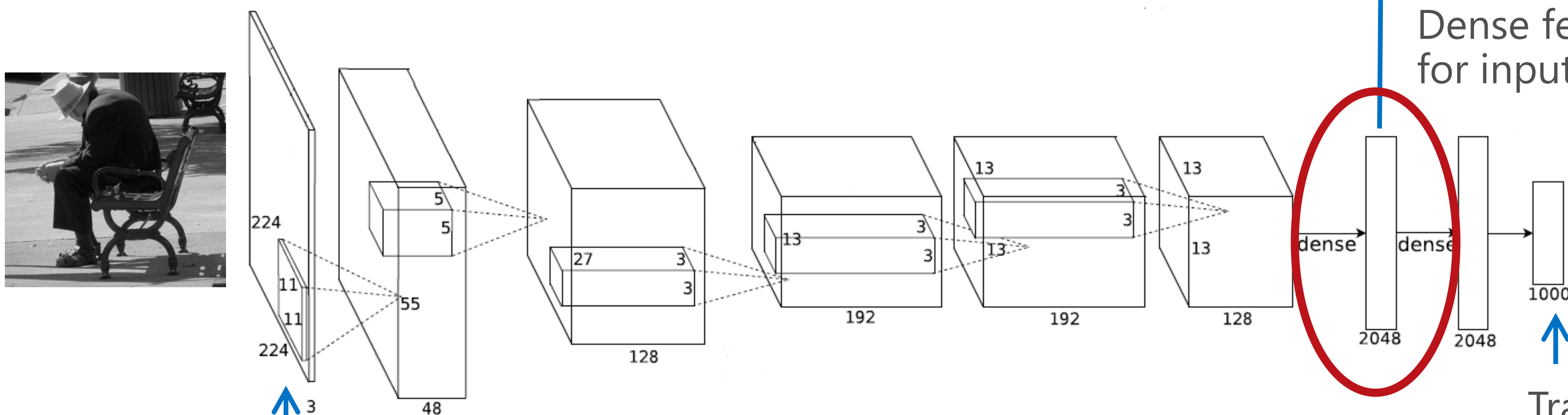
Huang, He, Gao, Deng, Acero, Heck, "Learning Deep Structured Semantic Model for Web Search," CIKM, 2013

The convolutional network at the image side

Feed the pre-trained image feature vector into the image side of the DMSM



Dense feature vector for input image



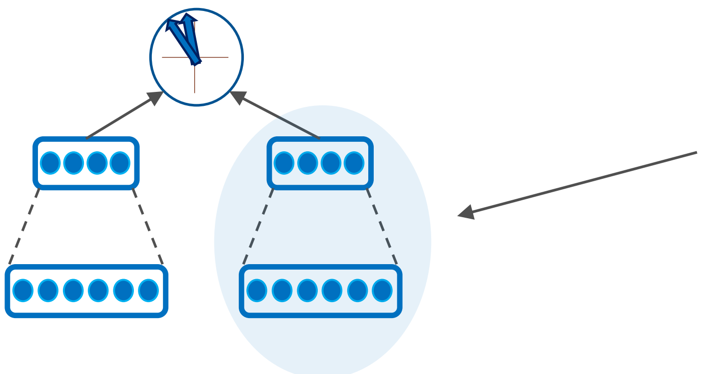
Raw pixels from input box

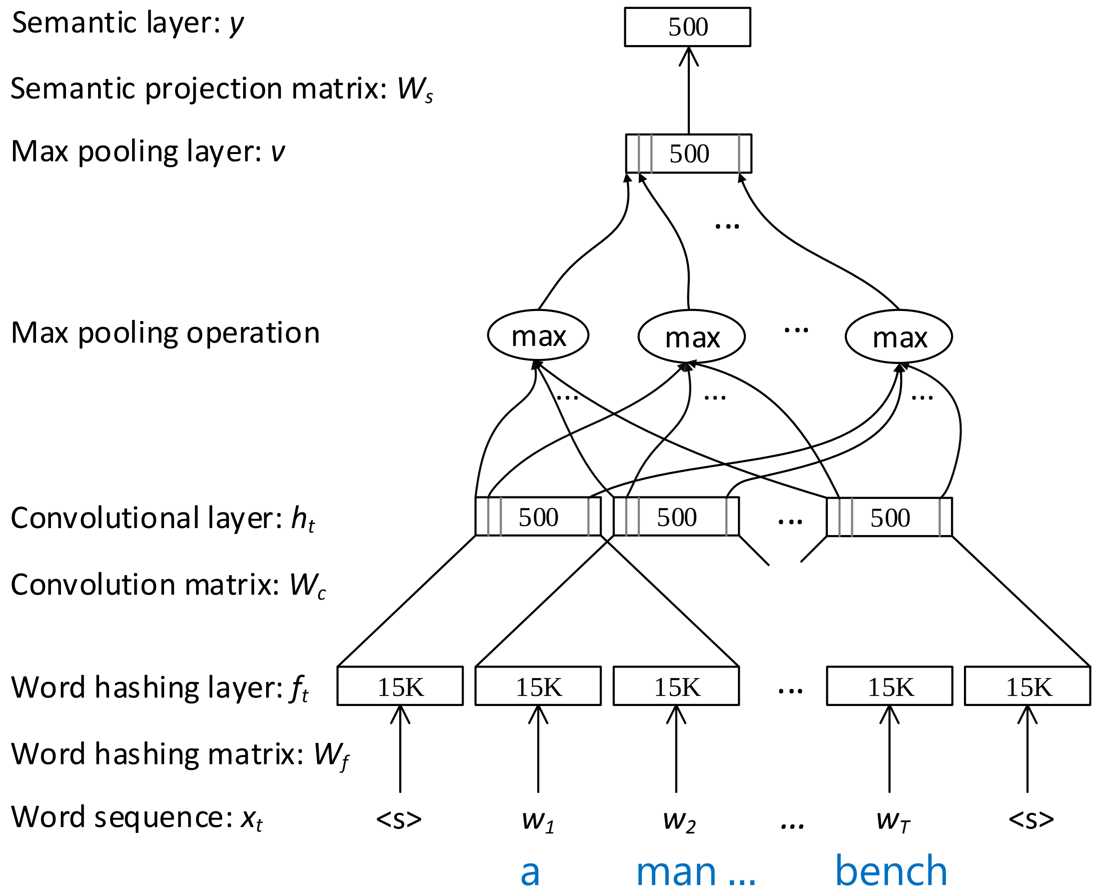
Trained to predict object in image

Tuned image features from AlexNet (Krizhevsky et al., 2012) or VGG (Simonyan and Zisserman, 2014).

The convolutional network at the caption side

Models fine-grained structural language information in the caption





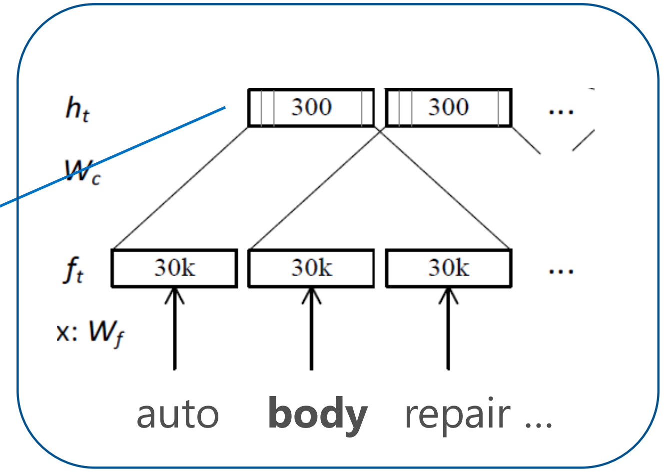
Using a convolutional neural network for the text caption side

Figure Credit: [Shen, He, Gao, Deng, Mesnil, WWW, April 2014]

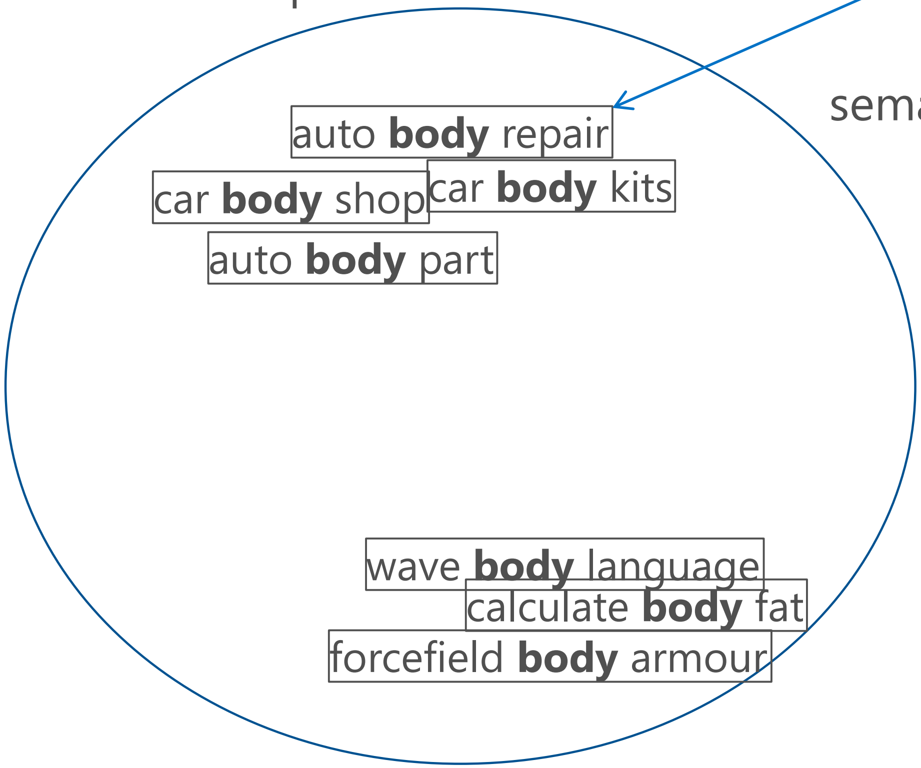
– What does the model learn at the convolutional layer?

Capture the local context dependent word sense

- Learn one embedding vector for each local context-dependent word



$$h_t = W_c \times [f_{t-1}, f_t, f_{t+1}]$$



semantic space

The similarity between different "***body***" within contexts

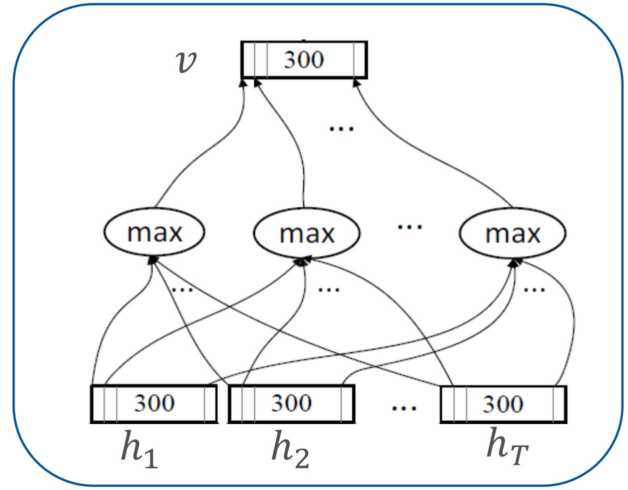
car **body** shop	**cosine similarity**
car **body** kits	0.698
auto **body** repair	0.578
auto **body** parts	0.555
wave **body** language	0.301
calculate **body** fat	0.220
forcefield **body** armour	0.165

high similarity

low similarity

What happens at the max-pooling layer?

- Aggregate *local topics* to form the *global intent*
- Identify salient words/phrase at the max-pooling layer



$$v(i) = \max_{t=1,\ldots,T}\{h_t(i)\}$$

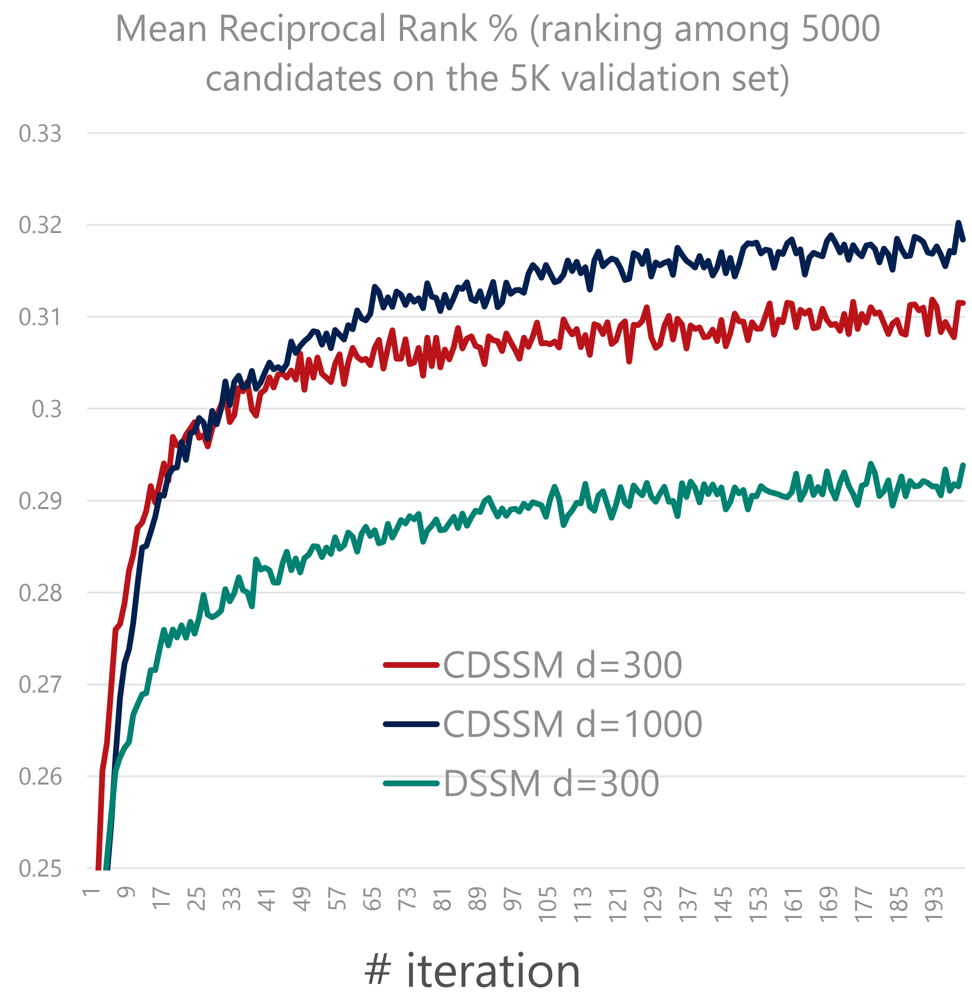
where $i = 1, \ldots, 300$

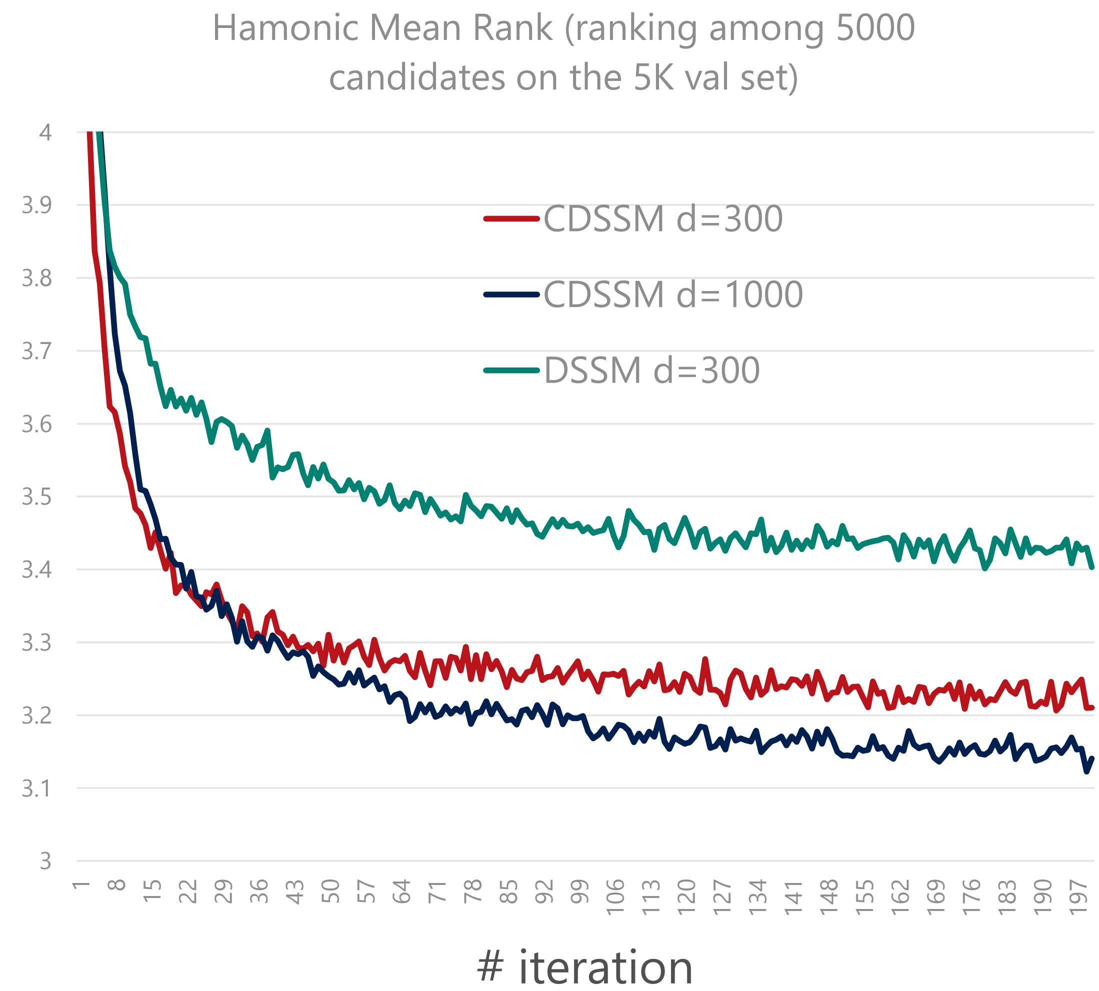
Words that win the most active neurons at the **max-pooling layers:**

Usually, those are salient words containing clear intents/topics

DMSM learning

Evaluation: on a 5K val set, for each image, rank the 5K captions and check the rank of the *true* caption





Evaluation

Human judgment is the ultimate metric

Turing Test Results

at the MS COCO Captioning Challenge 2015

Microsoft COCO
Common Objects in Context

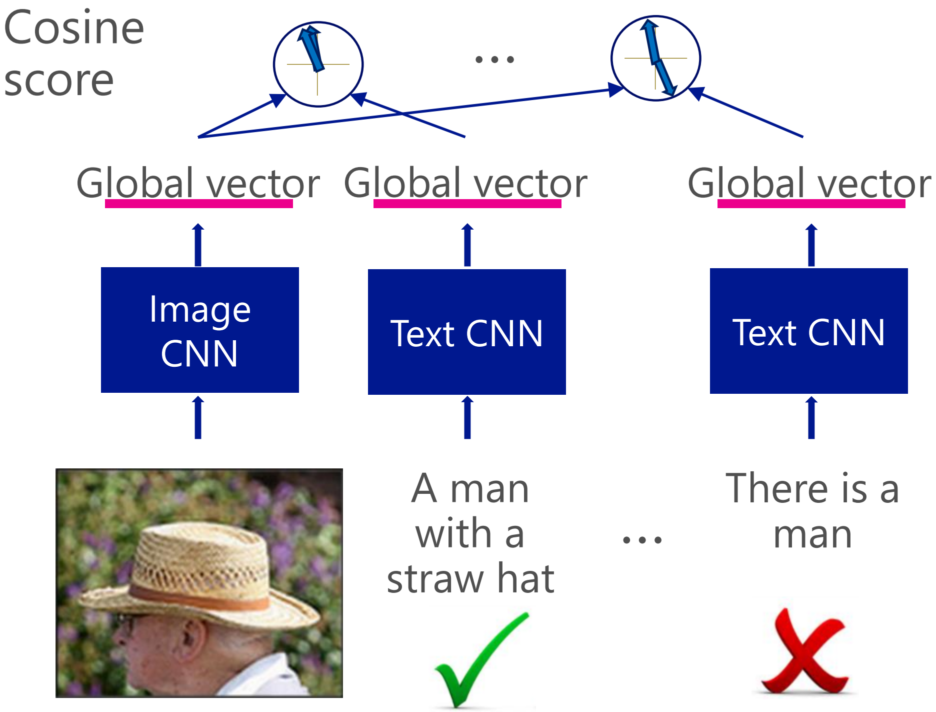
	% of captions that pass the Turing Test	Official Rank
MSR	32.2%	1st
Google	31.7%	1st
MSR Captivator	30.1%	3rd
Montreal/Toronto	27.2%	3rd
Berkeley LRCN	26.8%	5th
Other groups: Baidu/UCLA, Stanford, Tsinghua, etc.		
Human	**67.5%**	--

Still a big gap!

A brief comparison:

DMSM's objective:
the score of the reference to be higher than other generic captions.

MRNN's objective:
the score of the reference to be higher than arbitrary word sequences



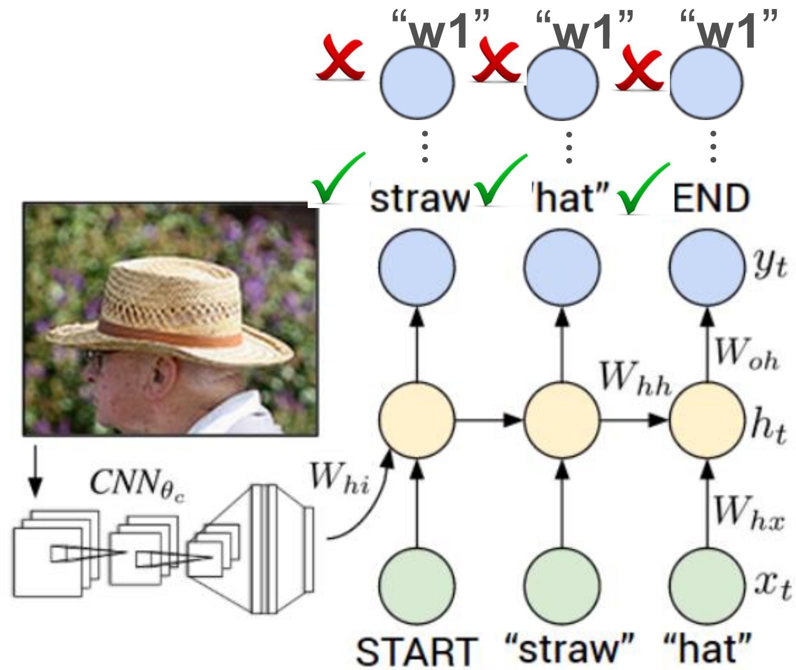


Image Credit: Karpathy and Fei-Fei 2015

DMSM focuses on semantics rather than syntax. E.g., ensures the reference (*semantically interesting*) scores higher than generic ones (grammatically correct but *semantically incorrect or boring*), while MRNN focus on syntax more.

Auto metric & Human Judge

- MELM+DMSM and MRNN obtain same BLEU score
- But humans prefer MELM+DMSM's output more

System		BLEU %	Better or Equal to Human
Model 1:	**MELM + DMSM**	25.7	34.0%
Model 2:	**MRNN**	25.7	29.0%

Human judgers shown generated caption and human caption, choose which is "better", or equal.

Devlin, Cheng, Fang, Gupta, Deng, He, Zweig, and Mitchell "Language Models for Image Captioning: The Quirks and What Works," ACL 2015

Image Diversity

- Test images bucketed based on visual overlap with training
  - MELM+DMSM does well on images with low overlap
  - MRNN does well on images with high overlap

Condition	Train/Test Visual Overlap BLEU		
	Whole Set	20% Least	20% Most
D-ME+DMSM	25.7	20.9	29.9
MRNN	25.7	18.8	32.0

BLEU scores

Rare images w.r.t. training set

Common images w.r.t. training set

Language Analysis

- MRNN weakness: Repeated captions
  - Table 1: MRNN repeat captions seen in training data verbatim more often
  - Table 2: Systems produce same captions multiple times; MRNN does it the most

Table 1: Percentage of Produced Testval Captions Found in Training Captions

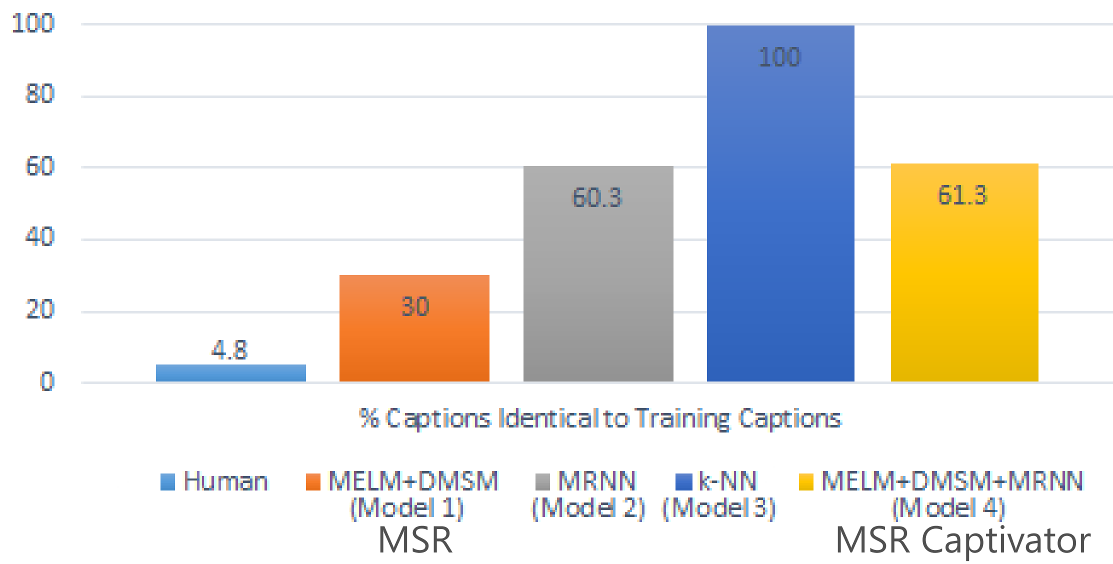
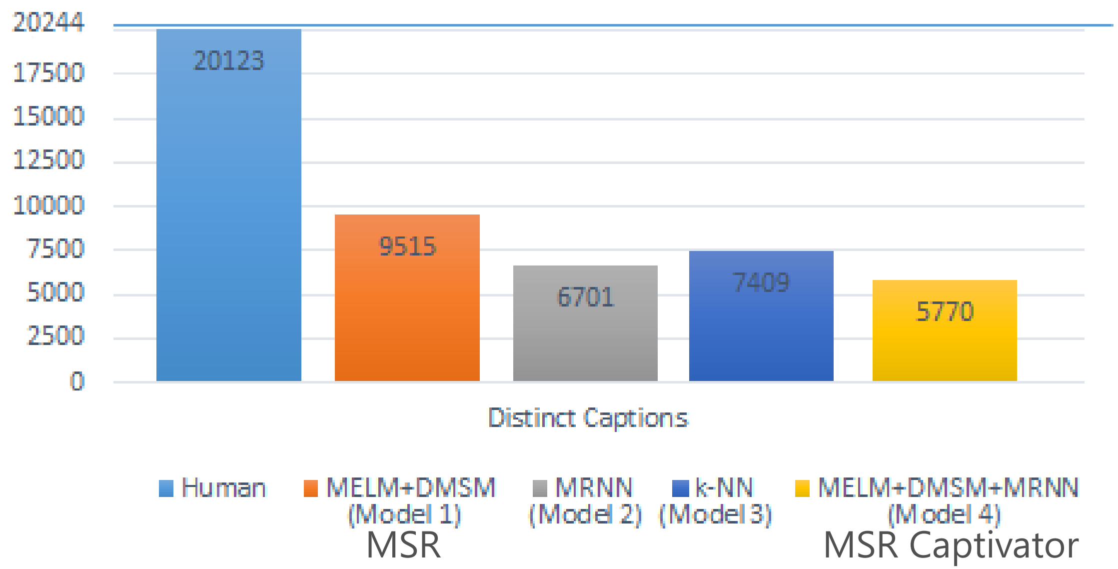


Table 2: Number *Distinct* Captions in Testval (out of 20,244 instances)



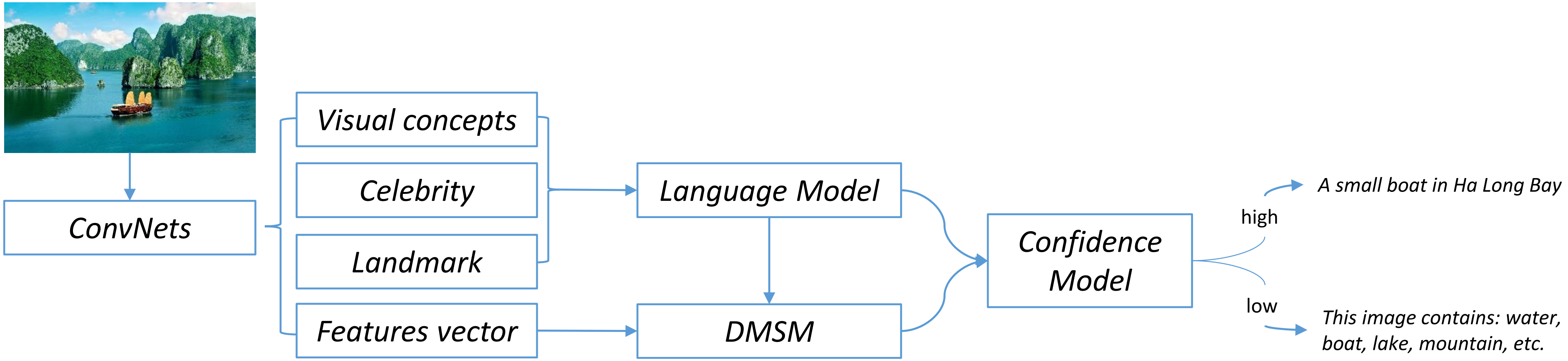
Example: MELM+DMSM: "A plate with a sandwich and a cup of coffee"

MRNN: "A close up of a plate of food" *(more generic)*

From COCO domain to open-domain

- Fast runtime
- Better accuracy per human judgment
- Broader coverage
- Richer information (e.g. people names, locations)
- Output with uncertainty information

Rich Image Captioning in the Wild



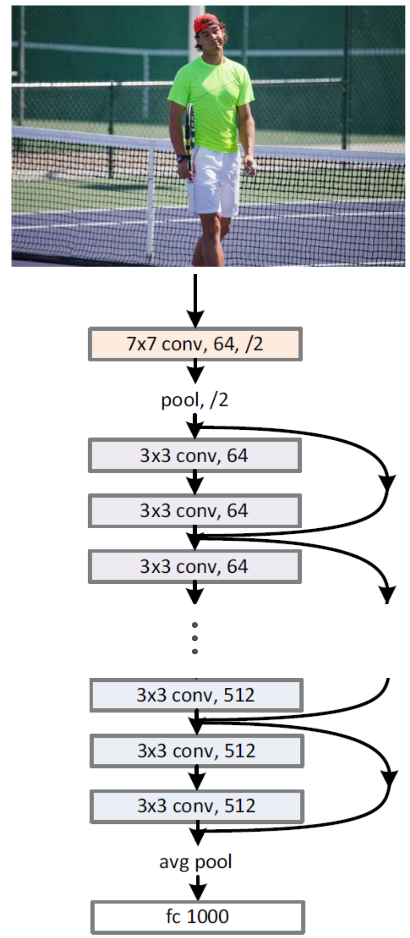
[Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, Chris Sienkiewicz submitted to CVPR Deep Vision 2016]

Deep ResNet for visual concepts detection

[He, Zhang, Ren, Sun, 2015]

ResNet
- ImageNet winning solution
- Treat as multiclass problem
- Sigmoid output
- No softmax normalization

Trained on multiple GPUs



man, tennis, court, holding, shirt, yellow, racquet, …

DMSM: Bridge the gap between image and language!

[Fang, et al., CVPR 2015]

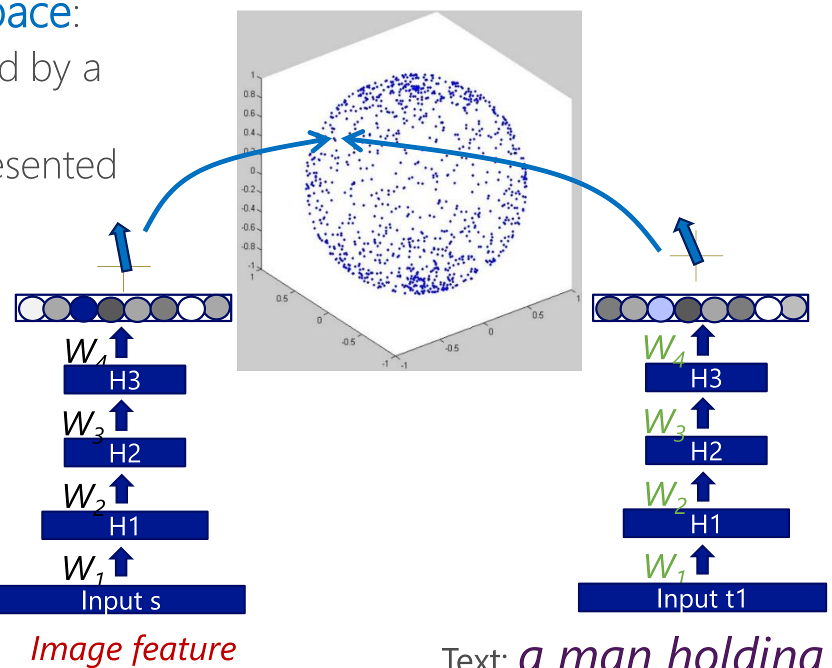
The deep multimodal semantic model projects images and captions to a abstract semantic space:

The overall semantics of a image will be represented by a vector in this space.

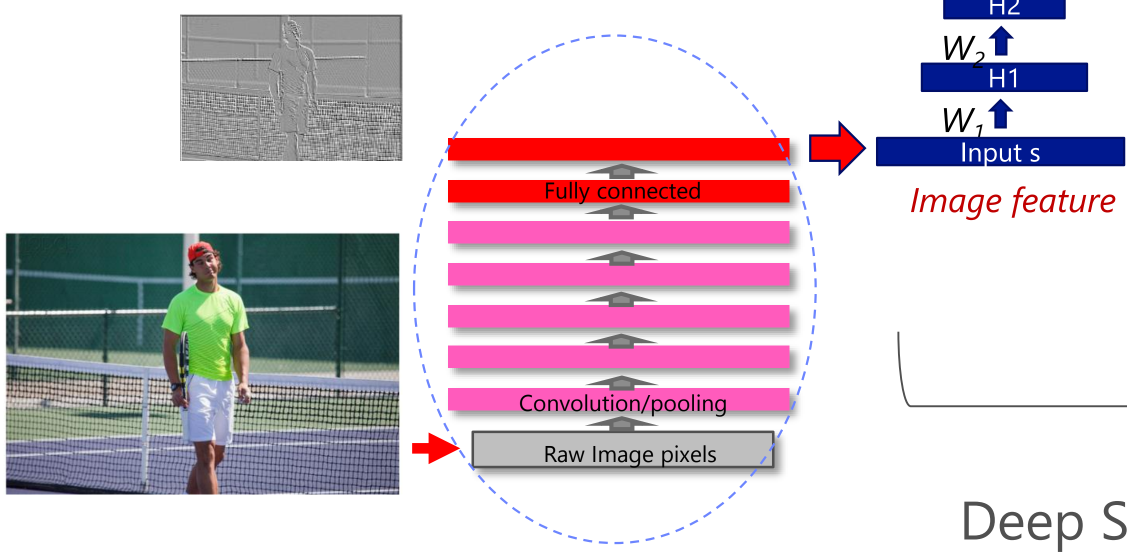
The overall semantics of a caption will also be represented by a vector in this space.

If these two vectors are close to each other, then the caption is a good match for the image.

Otherwise, not a matching caption.



Text: *a man holding a tennis racquet on a tennis court*



Deep Structured Semantic Model

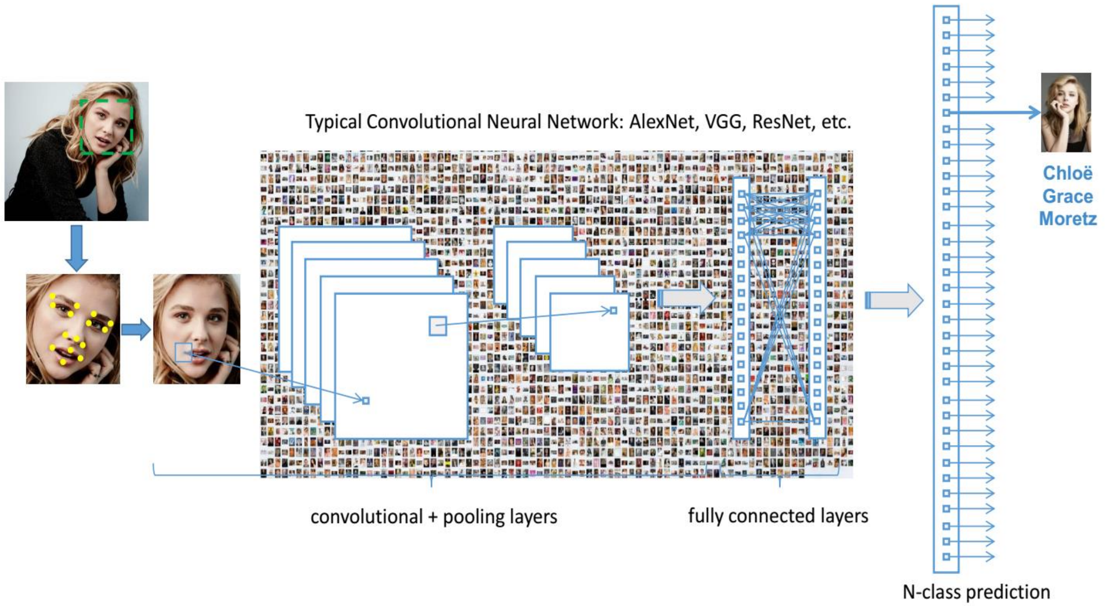
[Huang, He, Gao, Deng et al., 2013]

Deep Residual Network

[He, Zhang, Ren, Sun, 2015]

Entity Recognition

- Extreme classification with a **big** set of celebrities

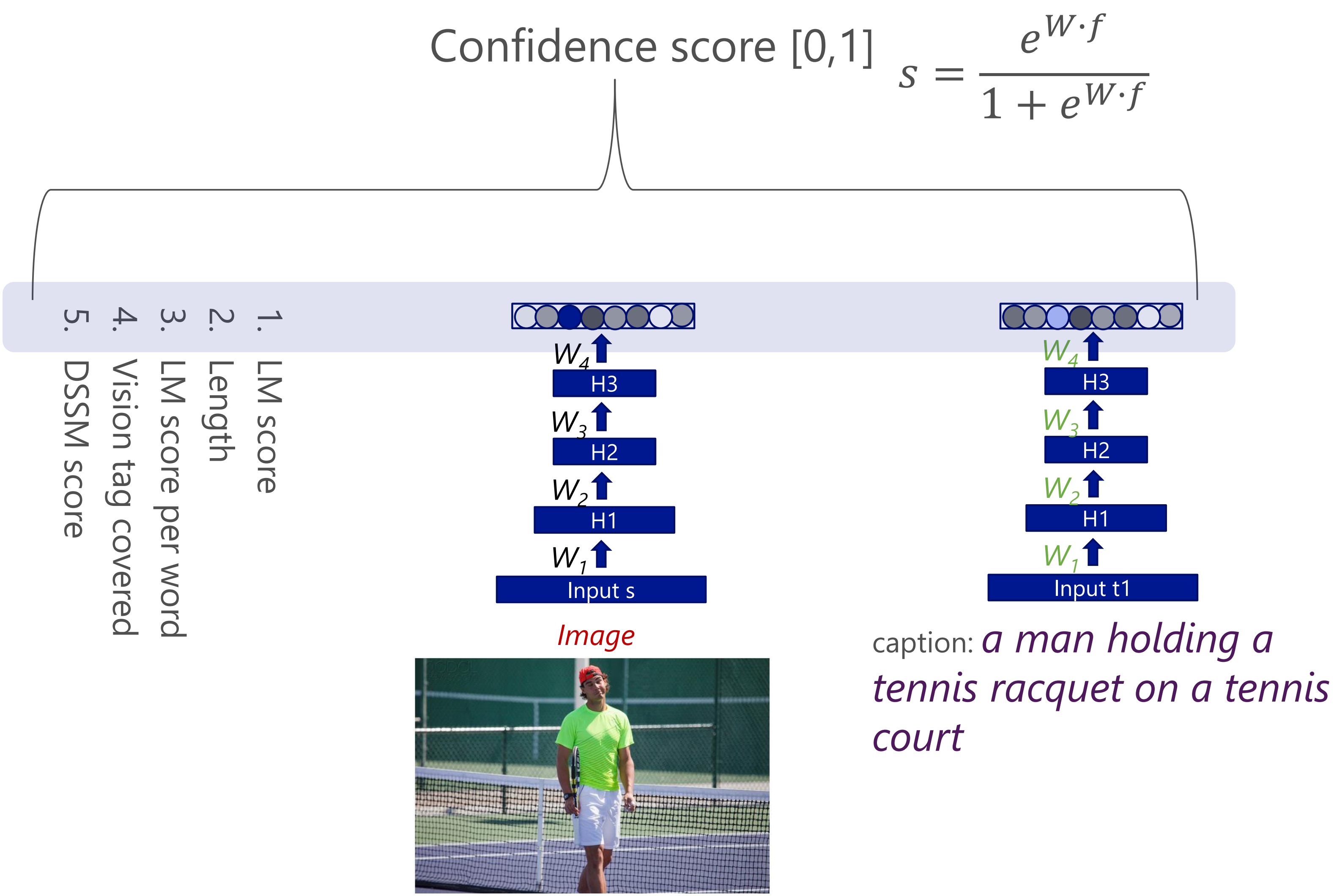


"Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background."

- Integrating entities (celebrities, landmarks, etc.) makes captions much richer.

[Guo, Zhang, Hu, He, Gao, MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World, 2016]

Describe with uncertainty



Test results - COCO

Beat previous SOTA on in-domain data (MS COCO)

System	Excellent	Good	Bad	Embarrassing
Fang et al., 2015	40.6%	26.8%	28.8%	3.8%
New system	51.8%	23.4%	22.5%	2.4%

Human evaluation on 1000 random samples of the COCO test set.

Test results - Instagram

Significantly beat previous SOTA on data in the wild

System	Excellent	Good	Bad	Embarrassing
Fang et al., 2015	12.0%	13.4%	63.0%	11.6%
New system	25.4%	24.1%	45.3%	5.2%

Human evaluation on Instagram test set, which contains 1380 random images that we scraped from Instagram.

Confidence score distribution - Instagram

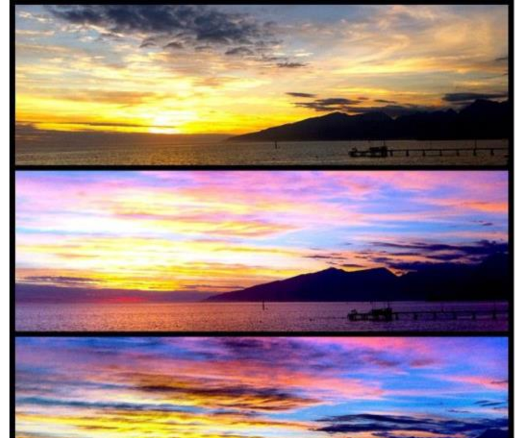
Confidence score aligns with human judgement well

Conf. score	Excellent	Good	Bad	Embarrassing
mean	0.59	0.51	0.26	0.20
Std dev	0.21	0.23	0.21	0.19

Above: Fang2015
Below: Ours

a dog sitting on top of a grass covered field
a dog sitting in the grass

a man holding a baseball bat at a ball
a man swinging a baseball bat in front of a crowd



a view of a sunset over water
a view of a sunset in the ocean

a black and white photo of a man wearing a hat
a man wearing a bow tie looking at the camera

a man wearing a suit and tie
Ian Somerhalder wearing a suit and tie

a man on a skateboard
this picture is about photo

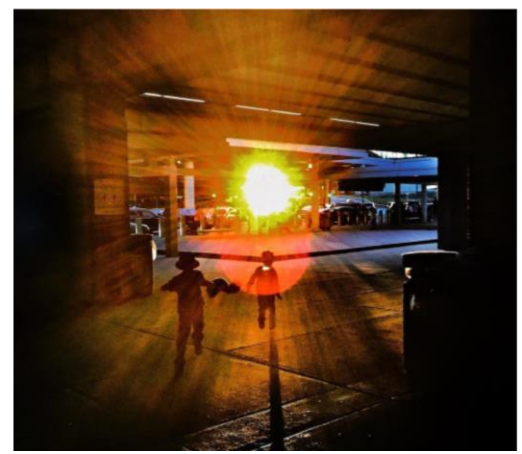
a man taking a picture in front of a mirror
an picture about person

a man holding a stop sign
a man holding a stop sign

a woman standing in front of a christmas tree
a woman standing next to a window

a colorful kite flying in the air
a table topped with a kite

a black and white photo of a man wearing a hat
a man posing for a picture



a couple of people at night
a fire hydrant that is lit up at night

a woman sitting on a couch
this picture is about person

a pair of scissors sitting on top of a table
a bunch of different items

a woman holding a red umbrella
the image is about person

a group of pictures on the wall
this picture seems contain text

two women standing in front of a cake
a woman posing for a picture

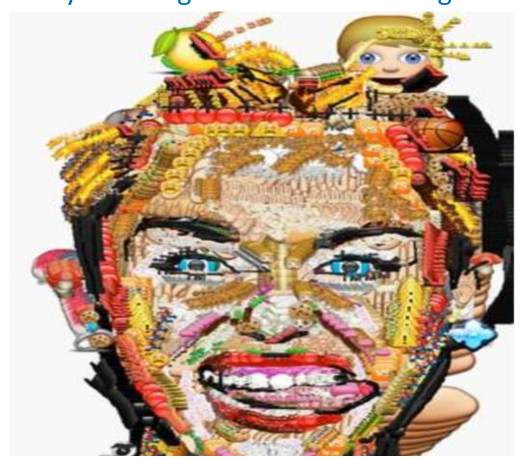
a woman sitting on a bench
a woman sitting on a bench

a man holding a baseball bat on a field
a boy standing in front of a building

a black and white photo of a woman brushing her hair
a woman standing in front of a mirror

a person holding a cell phone
a hand holding a cell phone

a man and a woman wearing a tie
a couple posing for a photo

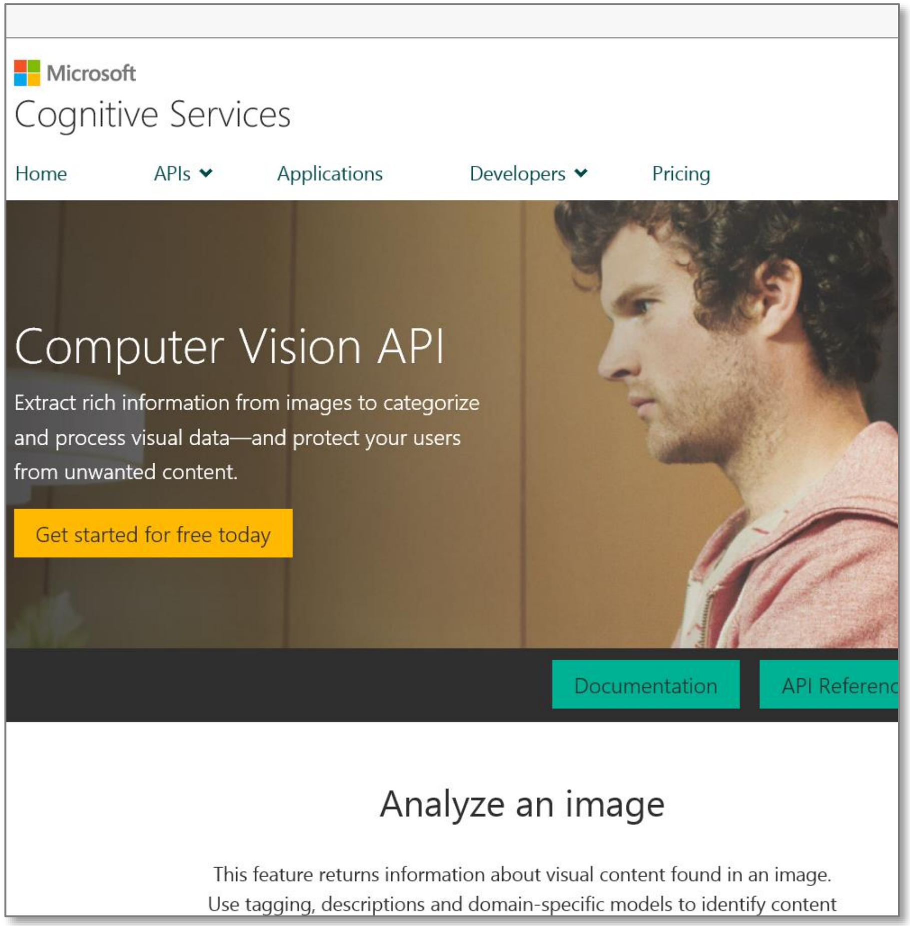


a man holding a teddy bear
a picture about table

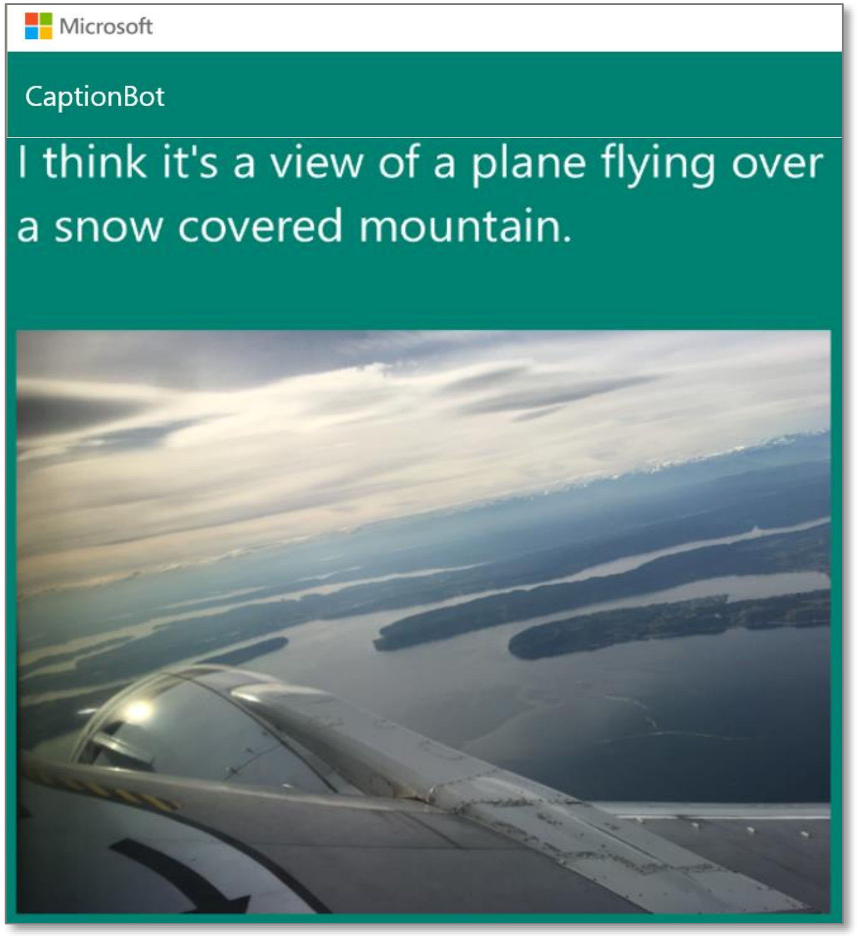
a pair of scissors
the image is about clothing

Image Captioning as a Cloud Service

API Accessible via MS Cognitive Services



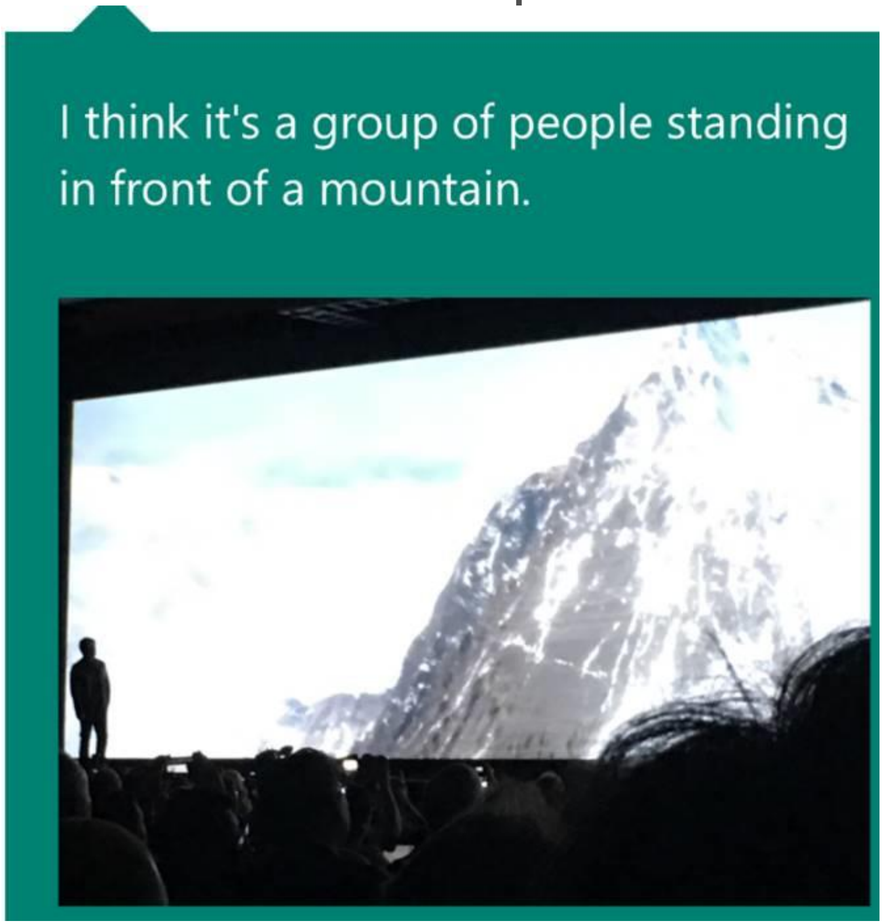
http://CaptionBot.ai



Example

I am not really confident, but I think it's Leonardo da Vinci sitting in front of a mirror and she seems 😐.

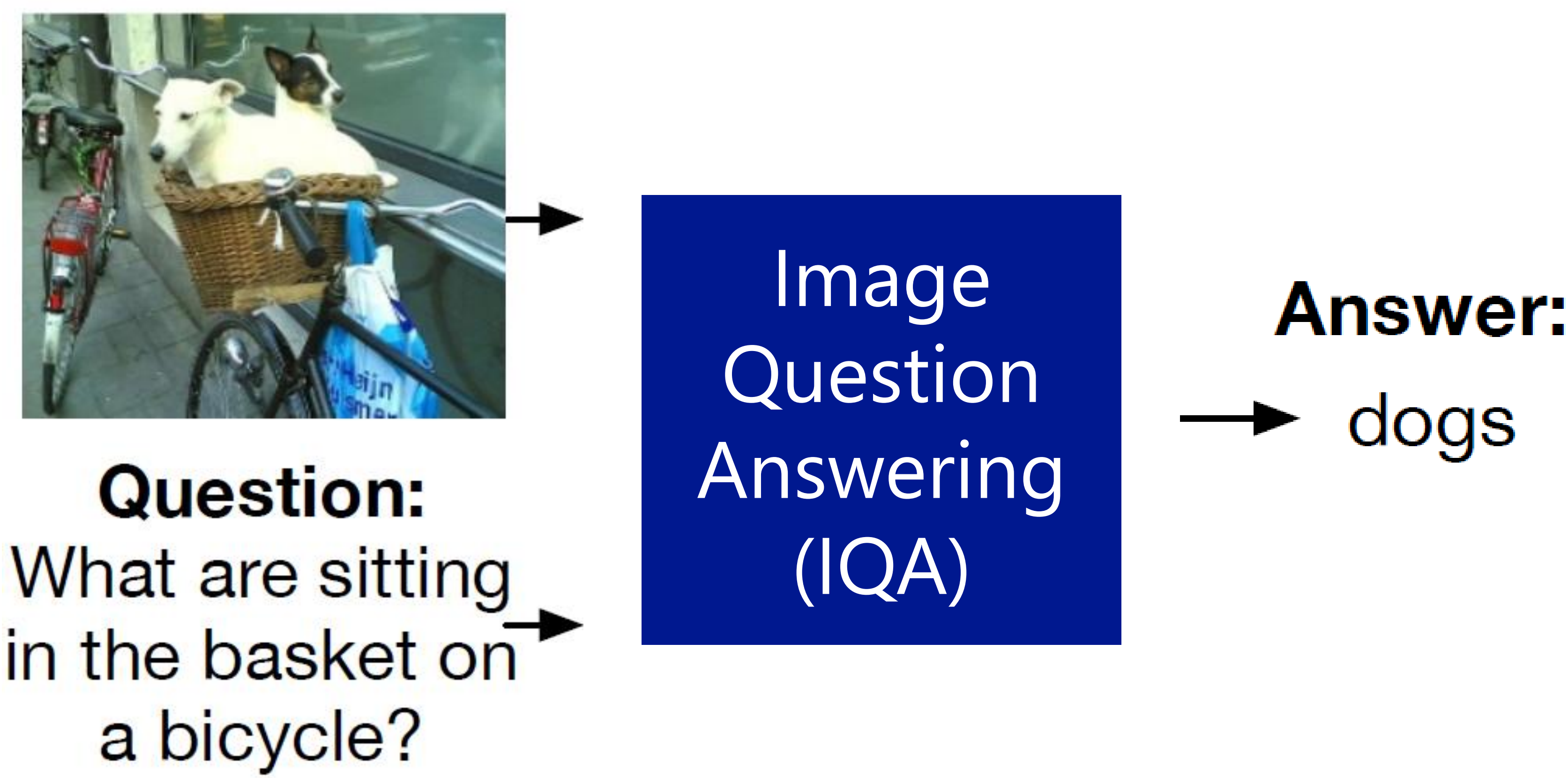
More examples from GTC



when Jen-Hsun Huang was giving a keynote showing off a GPU-powered VR visiting of mt. Everest -- here is what our CaptionBot has to say.

From Captioning to Question Answering

- Answer natural language questions according to the content of a reference image.



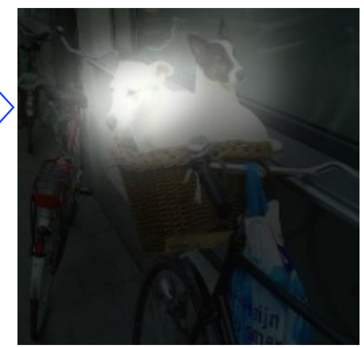
Caption vs. QA: need reasoning

Image QA: reasoning is a key.

Question:
What are sitting in the basket on a bicycle?

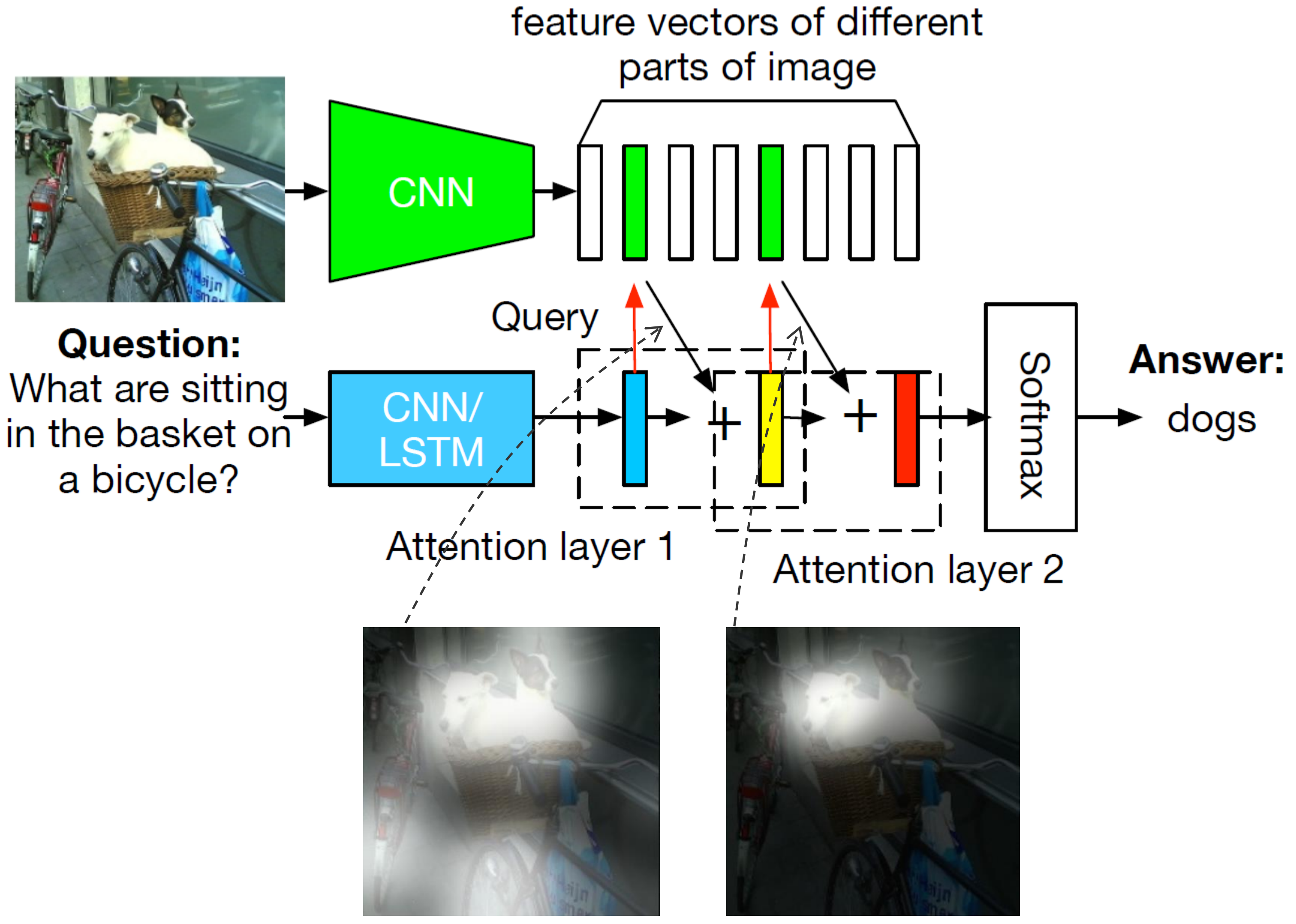
Multiple-steps of reasoning over the image to infer the answer

Answer:
dogs



Stacked Attention Network for Reasoning

IQA: Need perform multiple steps of reasoning over the image to infer the answer.



Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola, "Stacked Attention Networks for Image Question Answering," CVPR 2016 (oral)

Components of the SAN

- Image Model

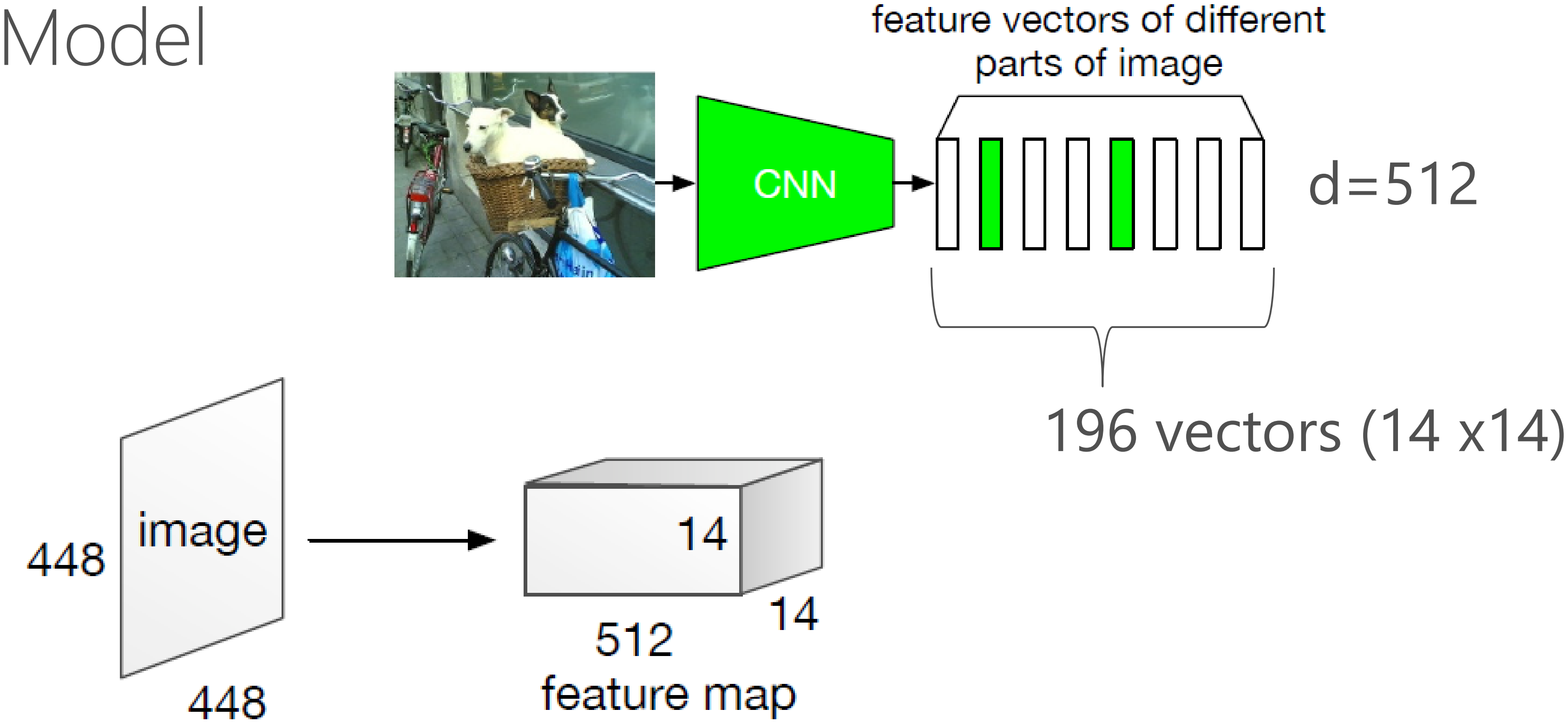


Figure 2: CNN based image model

$$f_I = \mathrm{CNN}_{vgg}(I).$$

Components of the SAN

- Question Model

  Code the question into a vector using LSTM

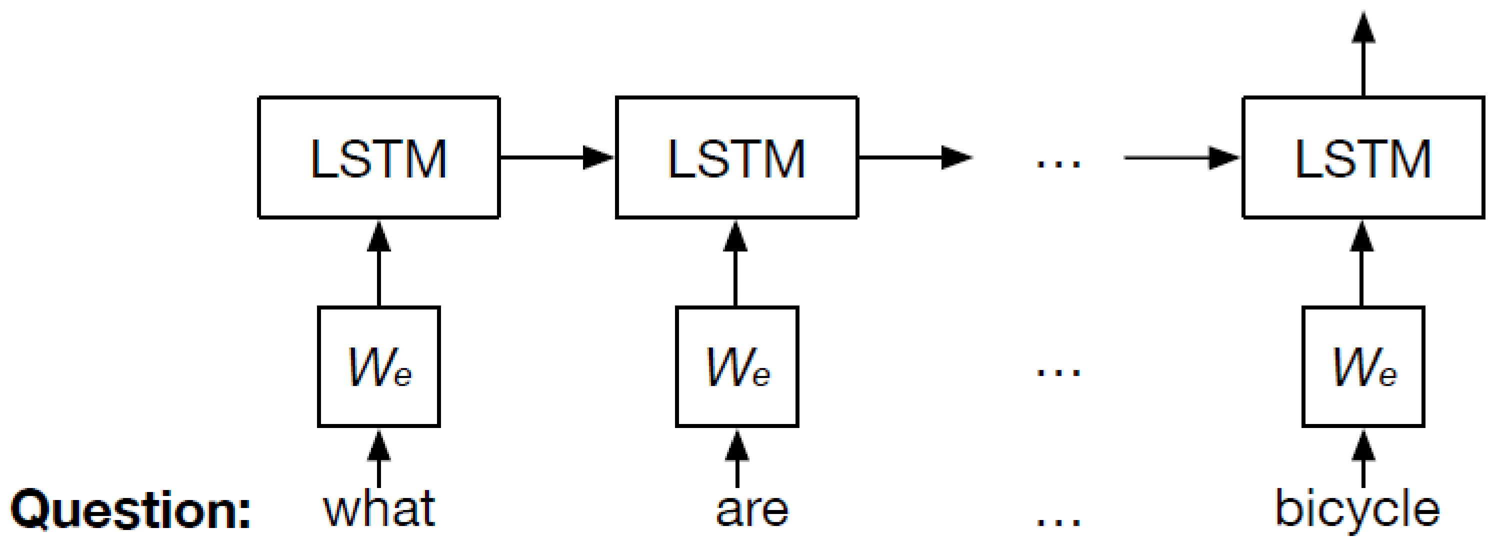
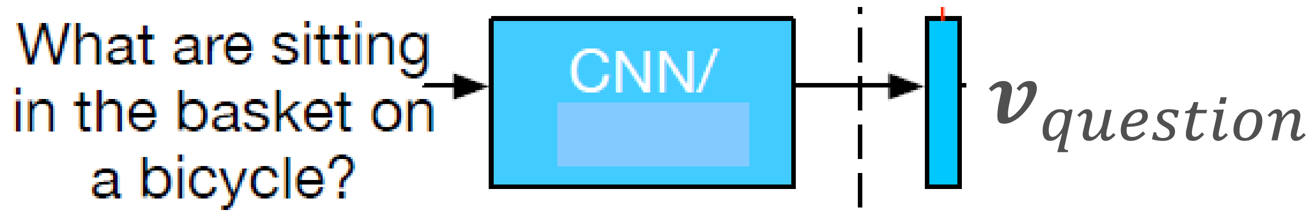


Figure 3: LSTM based question model

Components of the SAN

- Question Model

  Code the question into a vector using CNN



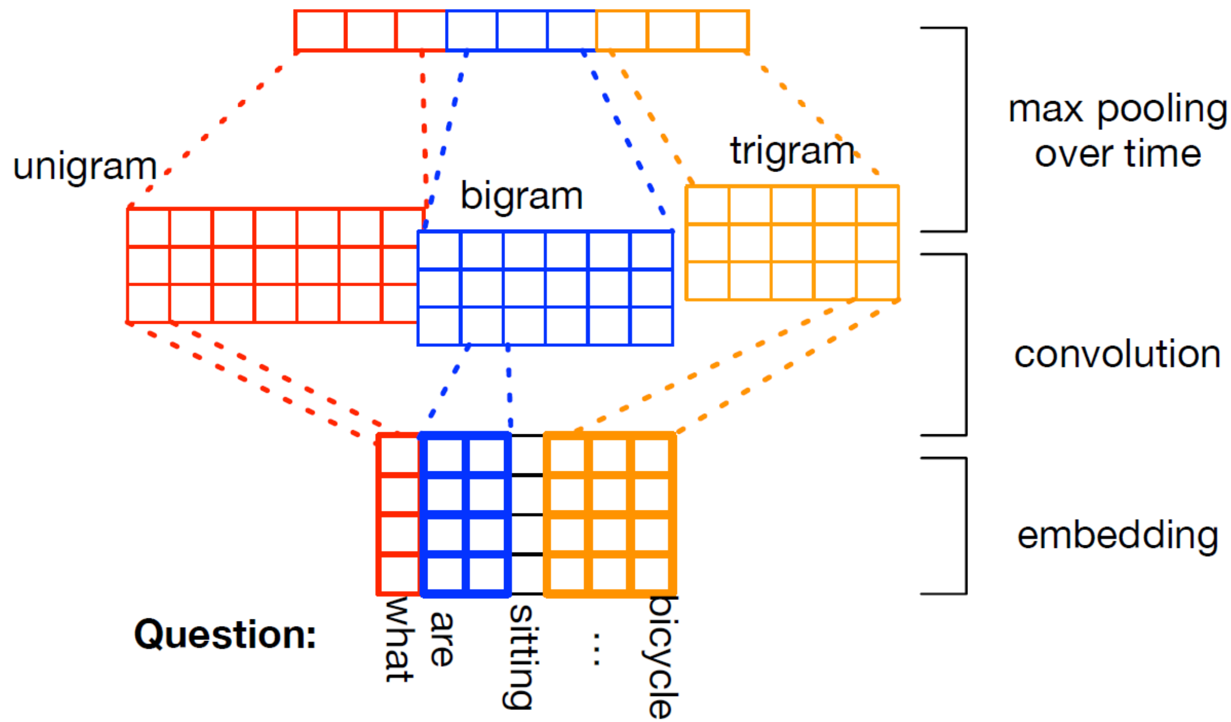
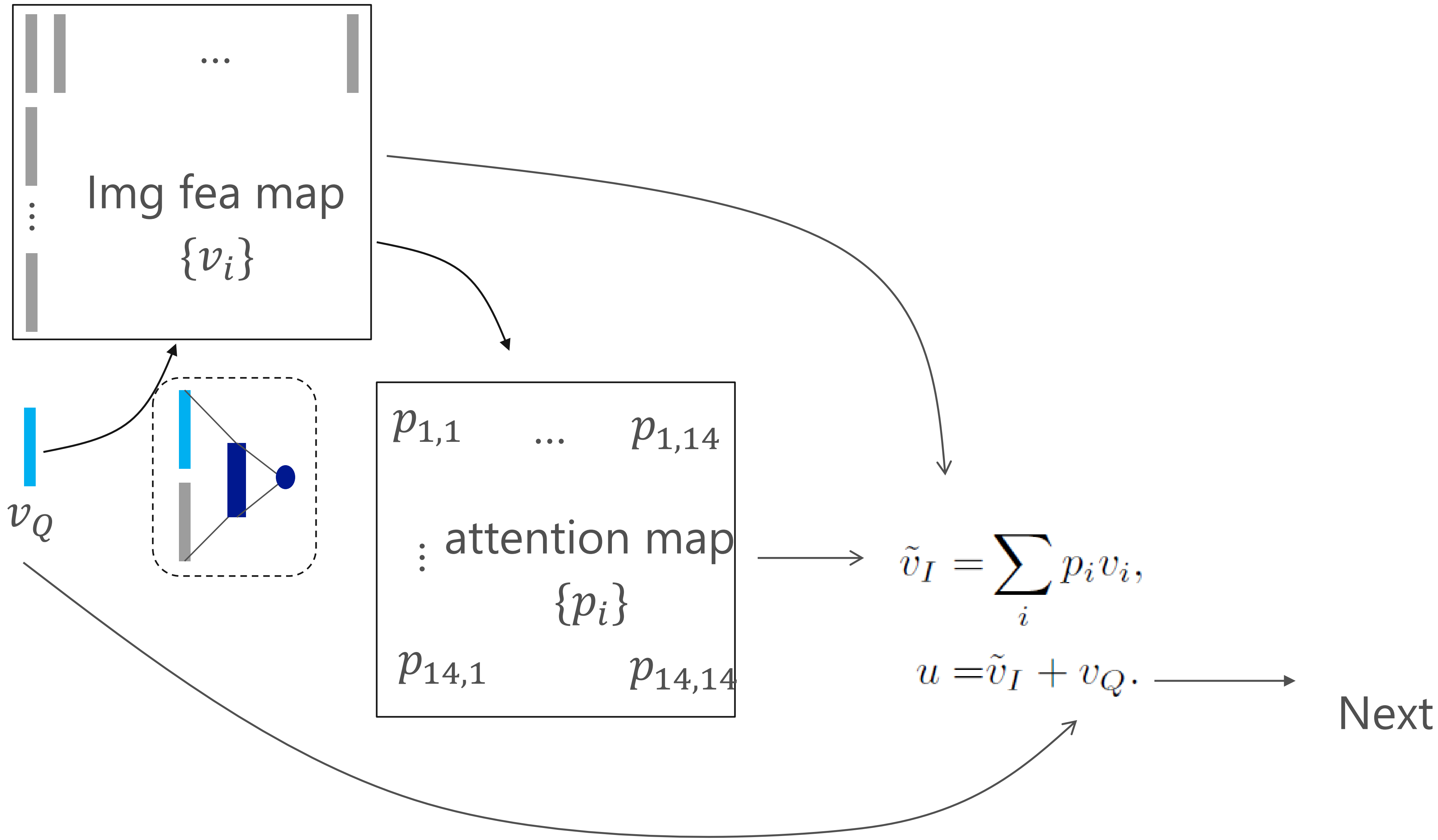
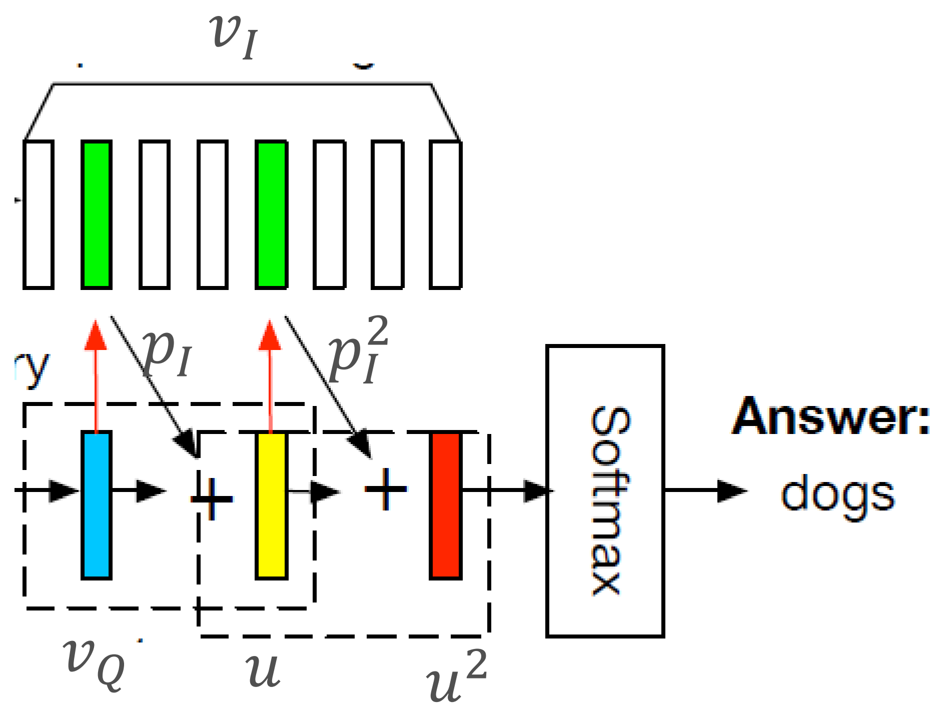


Illustration of computing the attention layer



Components of the SAN

- Stacked Attention Mechanism



1st attention layer:

$$h_A = \tanh(W_{I,A} v_I \oplus (W_{Q,A} v_Q + b_A)),$$
$$p_I = \text{softmax}(W_P h_A + b_P),$$

Deeper attention layers ($k = 2, \ldots$):

$$\tilde{v}_I = \sum_i p_i v_i,$$
$$u = \tilde{v}_I + v_Q.$$

$$h_A^k = \tanh(W_{I,A}^k v_I \oplus (W_{Q,A}^k u^{k-1} + b_A^k)),$$
$$p_I^k = \text{softmax}(W_P^k h_A^k + b_P^k).$$
$$\tilde{v}_I^k = \sum_i p_i^k v_i,$$
$$u^k = \tilde{v}_I^k + u^{k-1}.$$

$$p_{\text{ans}} = \text{softmax}(W_u u^K + b_u).$$

Results

	test-dev				test-std
Methods	All	Yes/No	Number	Other	All
VQA: [1]					
Question	48.1	75.7	36.7	27.1	-
Image	28.1	64.0	0.4	3.8	-
Q+I	52.6	75.6	33.7	37.4	-
LSTM Q	48.8	78.2	35.7	26.6	-
LSTM Q+I	53.7	78.9	35.2	36.4	54.1
SAN(2, CNN)	**58.7**	**79.3**	**36.6**	**46.1**	**58.9**

Table 5: VQA results on the official server, in percentage

Other:
Object
Color
Location …

Big improvement on the VQA benchmark and other benchmarks

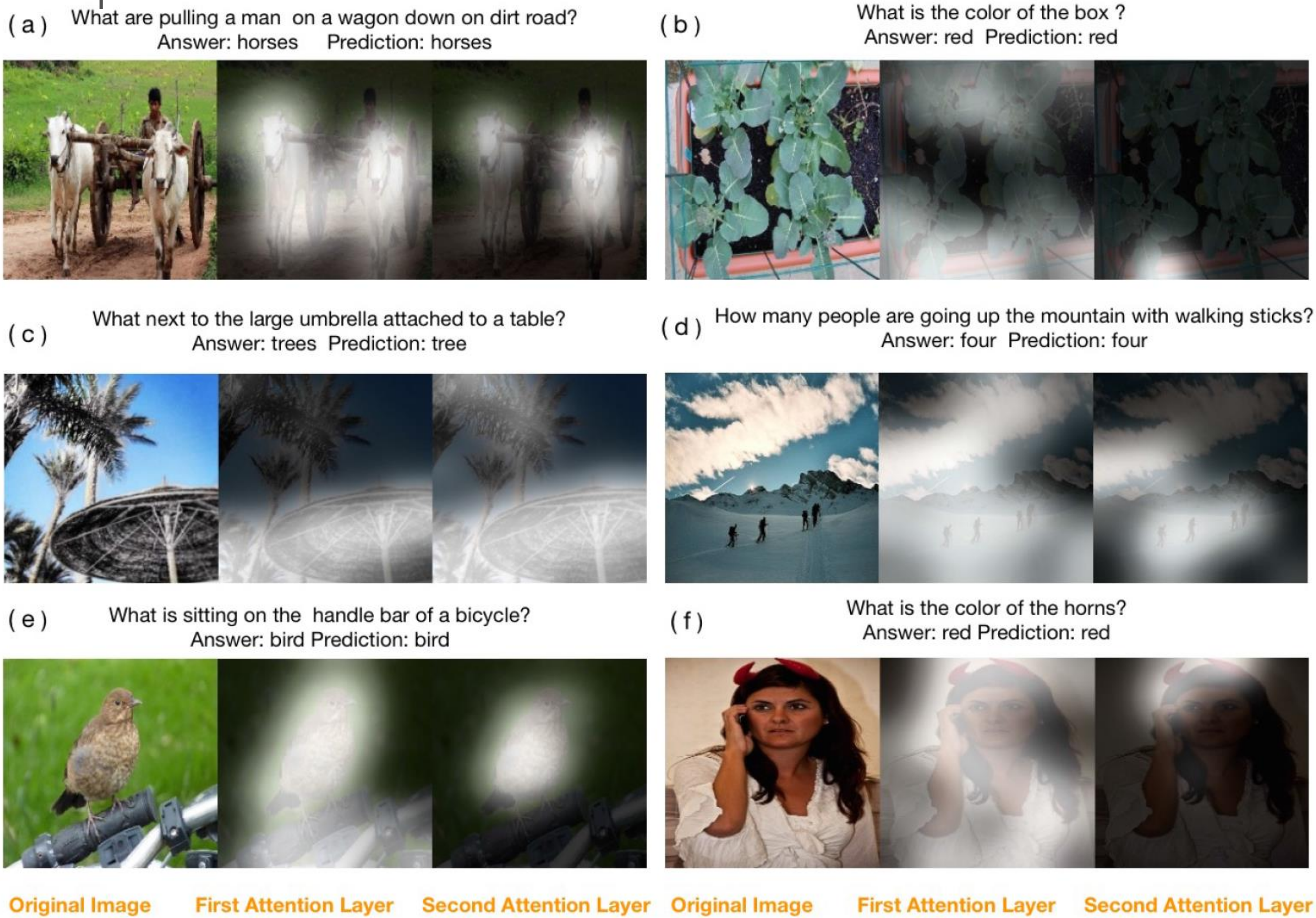
Q: what stands between two blue lounge chairs on an empty beach?

1st attention layer

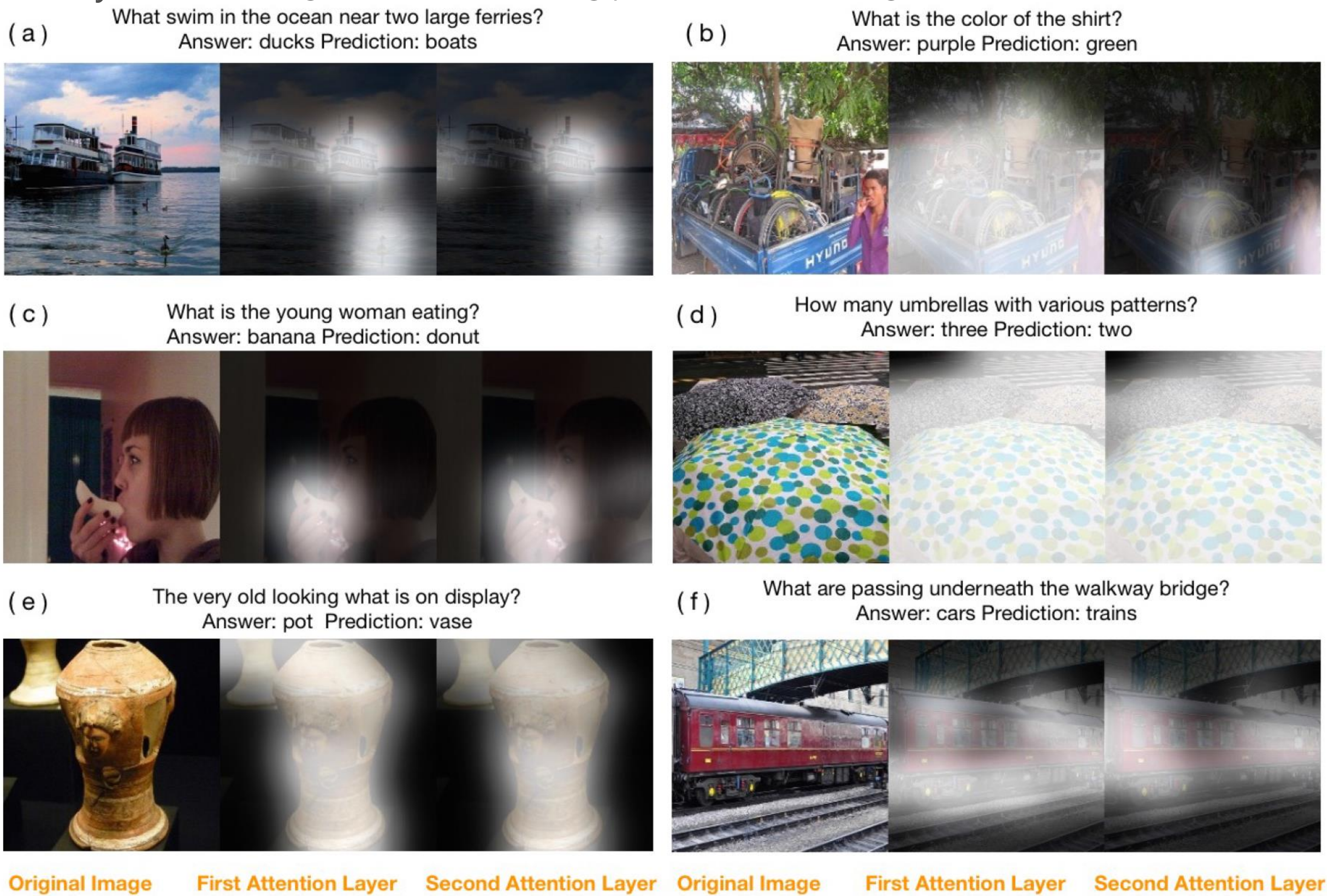
2nd attention layer

Ans: **umbrella**

More examples:



Error analysis: 22% wrong attention, 42% wrong prediction, 31% ambiguous answer, 5% label error



Go deeper?

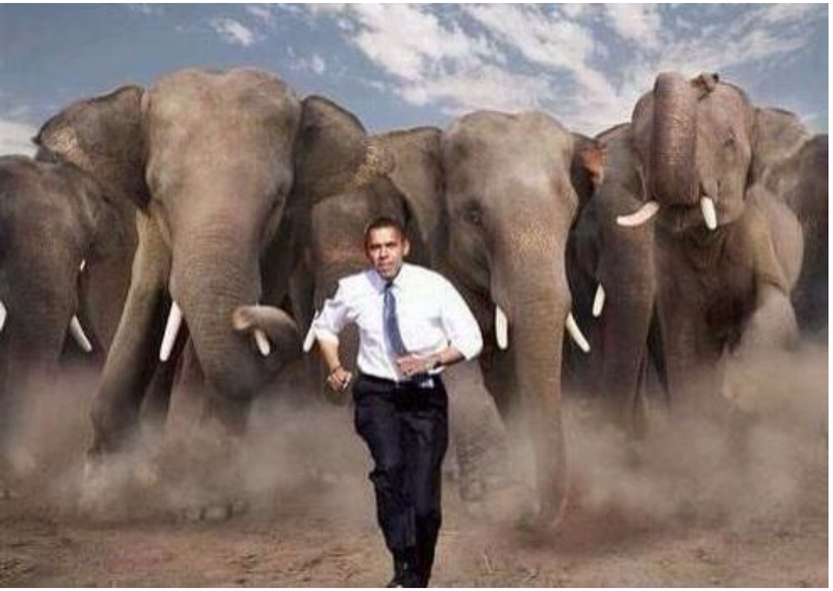


Image credit:
http://s122.photobucket.com/user/bmeuppls/media/stampede.jpg.html

before:

→ a herd of elephants standing next to a man

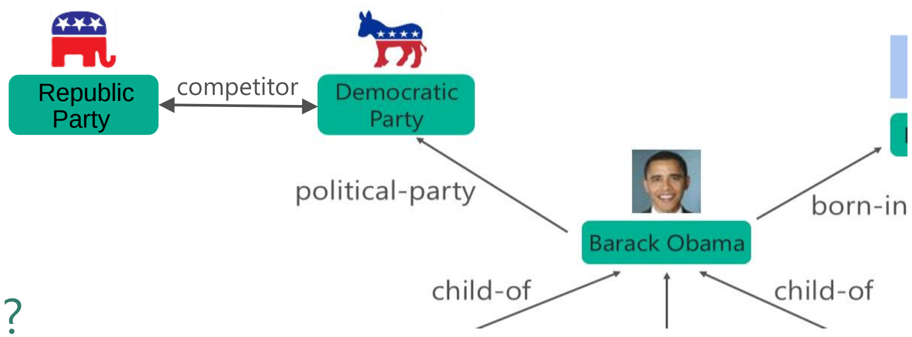
Now, + Entity:

→ a herd of elephants standing next to Obama

Next, + knowledge & reasoning:

Obama is *the president from* the **Democratic party**,
whose *competitor is* the **Republic party**,
whose *mascot is* **Elephant**.

→ Obama is chased by his republic competitors ☺



Knowledge Graph

Who is that person?

What are behind that man?

Why these elephants are chasing him?

Other relevant work

- Question answering/Inference
  - David Golub, Xiaodong He, "Character-Level Question Answering with Attention," arXiv 1604.00727
  - Moontae Lee, Xiaodong He, Wen-tau Yih, Jianfeng Gao, Li Deng, and Paul Smolensky, Reasoning in Vector Space: An Exploratory Study of Question Answering, ICLR 2016
  - Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, Mari Ostendorf, "Deep Reinforcement Learning with an Action Space Defined by Natural Language," arXiv:1511.04636
  - Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao, Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base, ACL 2015
  - Wen-tau Yih, Xiaodong He, and Christopher Meek, Semantic Parsing for Single-Relation Question Answering, ACL 2014

- Knowledge/Language representation
  - Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng, Embedding Entities and Relations for Learning and Inference in Knowledge Bases, ICLR 2015
  - Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck, Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, ACM International Conference on Information and Knowledge Management, 2013
  - Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang, Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval, in *NAACL* 2015

Summary

- *Language* is a valuable supervision for teaching machines to understand complex scenes *as humans do*.
- Deep learning models can perform certain level of *reasoning* in the image-language joint space and answer questions
- Need to add *knowledge* to give machines the common sense beyond in an isolated image

Please try out http://CaptionBot.ai ☺