Part V

Deep Reinforcement Learning for NLP

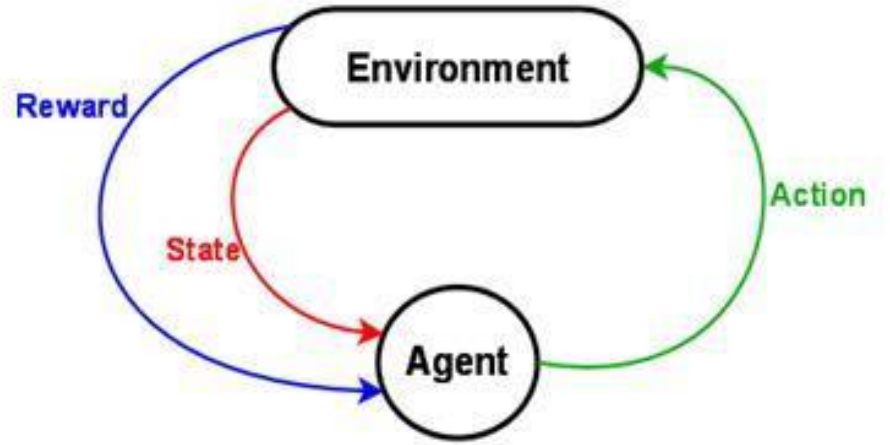
Deep reinforcement learning for NLP

- Deep RL background
- Deep RL for NLP

Background of reinforcement learning

Reinforcement learning model:

- environment state set: *S*
- Action set: *A*
- rules of transitioning between states
- rules that determine the immediate reward of a state transition
- rules that describe what the agent observes

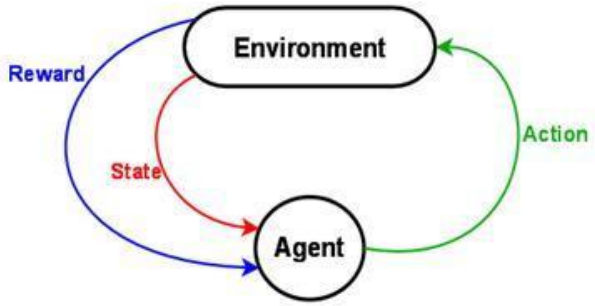


Sutton, Richard S.; Barto, Andrew G. (1998). Reinforcement Learning: An Introduction. MIT Press.

The Reinforcement Learning Problem

- Markov Decision Process (MDP)
  - State s_t, action a_t, reward r_t
  - Policy: p(action|state), e.g., $p(a_t|s_t)$
  - Objective: Find the best policy
  - Best: Maximize the long-term reward

Q-Learning



Used to learn the policy of RL

Policy: a rule that the agent should follow to select actions given the current state

Q-Learning: find optimal policy for Markov decision process (MDP).

Approach: learning an action-value function, a.k.a. Q-function, that computes the expected utility of taking an action in a state – after training converges.

$$Q^{\pi}(s,a) = \mathbb{E}\left\{\sum_{k=0}^{+\infty}\gamma^{k} r_{t+k} \middle| s_t = s, a_t = a\right\},$$

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \eta_t \cdot \left(r_t + \gamma \cdot \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right)$$

In runtime: $a_t^* = argmax_{\{a\}}\{Q(s_t,a)\}$

Watkins and Dayan, (1992), 'Q-learning.' Machine Learning.

Recent success

- Deep Q-Network (DQN)

1. Task: playing Atari games
2. RL setting: **huge state space**, e.g., raw image pixels from screen shots. But **small action space**, e.g., possible move of the joystick.
3. Model: using convolutional neural networks to compute $Q(s, a)$.
4. Results: achieve human level performance

action

state

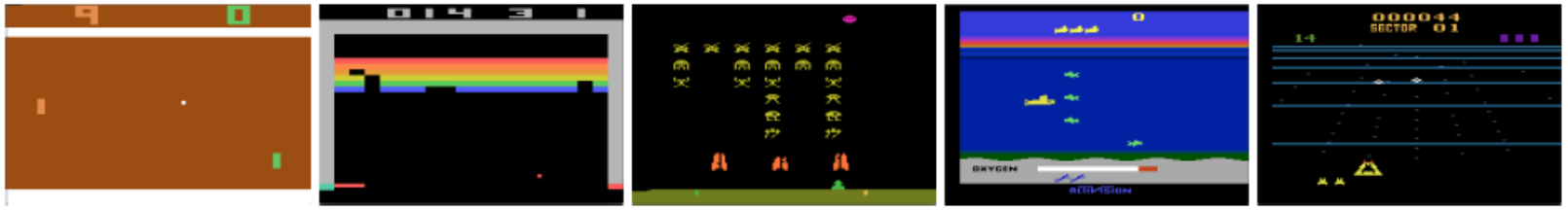


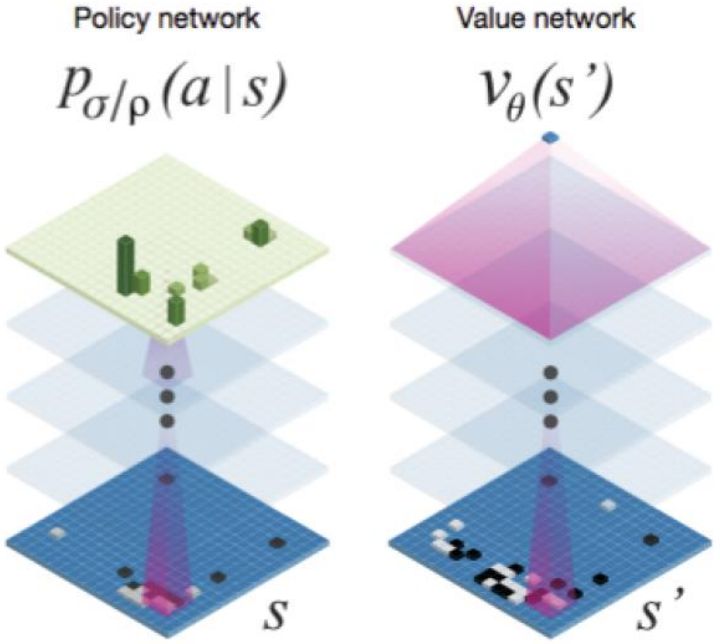
Figure 1: Screen shots from five Atari 2600 Games: (*Left-to-right*) Pong, Breakout, Space Invaders, Seaquest, Beam Rider

Mnih, Kavukcuoglu, Silver, Graves, Antonoglou, Wierstra, Riedmiller, "Playing Atari with Deep Reinforcement Learning", 2013

Recent success (cont.)

- AlphaGo

1. Task: playing Go
2. RL setting: **huge state space**, e.g., 19x19 board (highly complex and sensitive). But still **relatively small action space**, e.g., possible move (one out of <361 positions).
3. Model: built two deep networks: policy network and value network, both are CNNs
4. Use MCTS to look-ahead to estimate the value of states in a search tree.
5. Results: beat world Go Champion



Silver et al, "Mastering the game of Go with deep neural networks and tree search", 2016

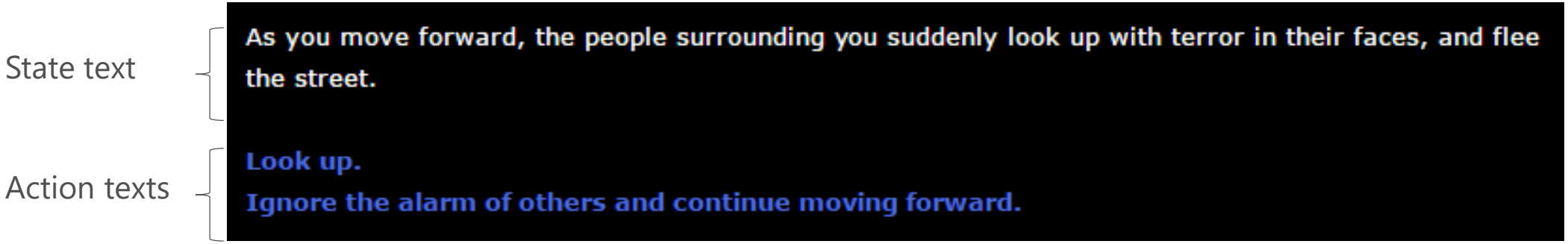
Reinforcement learning for language understanding

- Consider the sequential decision making problem for text understanding:
  - E.g., Conversation, Task completion, Playing text-based games…
  - At time t:
    - Agent observes the state as a string of text , e.g., state-text s_t
    - Agent also knows a set of possible actions, each is described as a string text, e.g., action-texts
    - Agent tries to understand the "state text" and all possible "action texts", and takes the **right** action – right means maximizing the long term reward
    - Then, the environment state transits to a new state, agent receives a immediate reward.

[Narasimhan, Kulkarni, Barzilay. EMNLP 2015]
[He, Chen, He, Gao, Li, Deng, Ostendorf, ACL 2016]
[He, Ostendorf, He, Chen, Gao, Li, Deng, EMNLP 2016]

RL with action space defined by NL

- Reinforcement learning (RL) with a natural language action space
  - Applications such as text games, webpage navigation, dialog systems (such as help desk and tutoring system)
  - Challenging because the potential action space is large and sparse
- Text-based game



Unbounded action space in RL for NLP

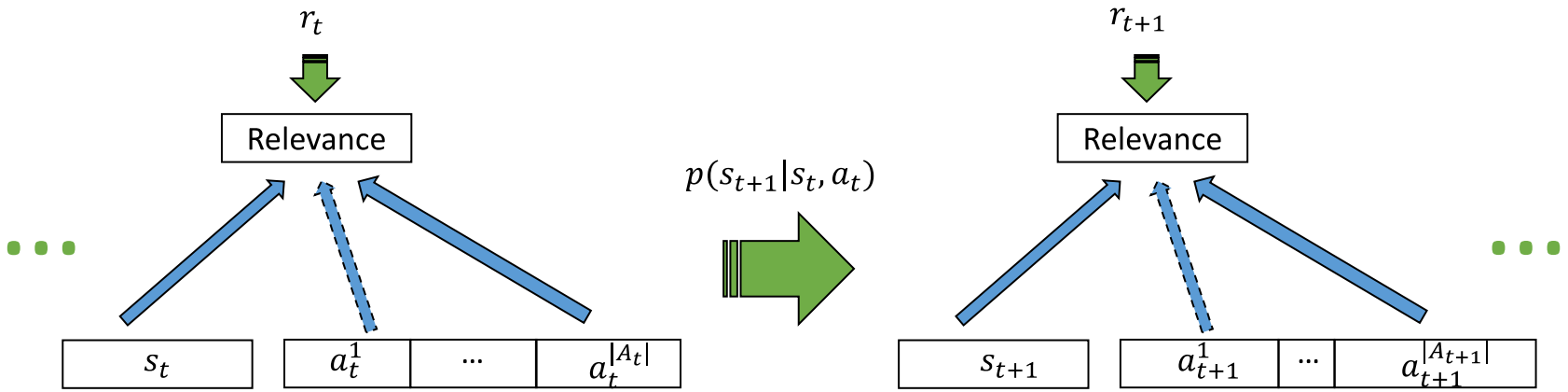
Not only the state space is huge, the action space is huge, too.

– Action is characterized by unbounded natural language description.

Example: a snapshot of a text-based game

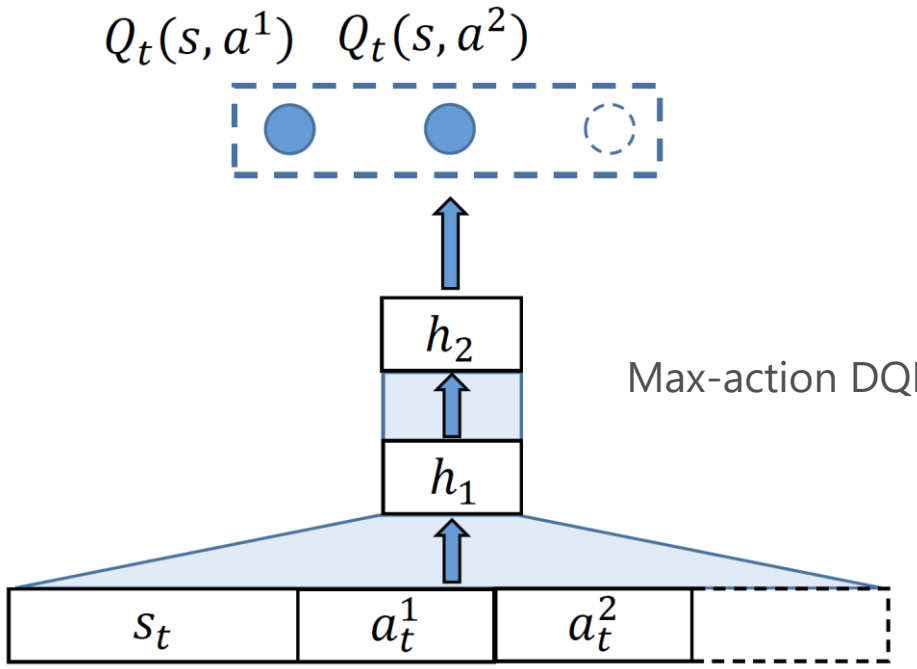
The Reinforcement Learning for NL problem

- A sequential text understanding problem
  - Unbounded state and action spaces (both in texts)
  - Time-varying feasible action set
    - At each time, the actions are different texts.
    - At each time, the number of actions are different.

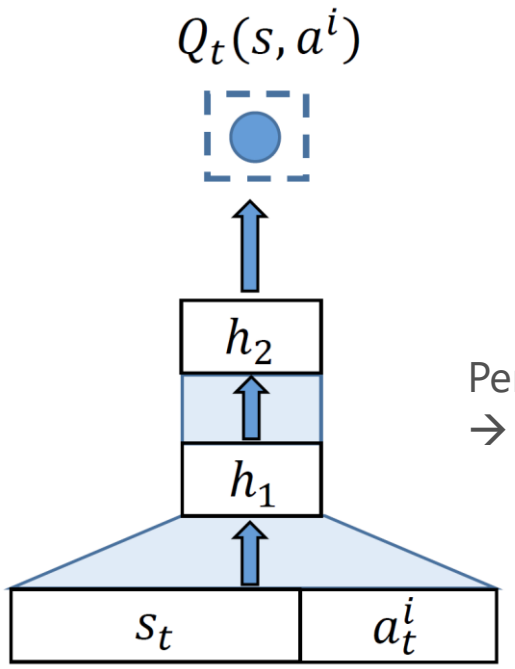


Baselines: Variants of Deep Q-Network

- Q-function: using a single deep neural network as function approximation
- Input: concatenated state-actions (BoW)
- Output: Q-values for different actions



Max-action DQN



Per-action DQN
→ max over a_t^i

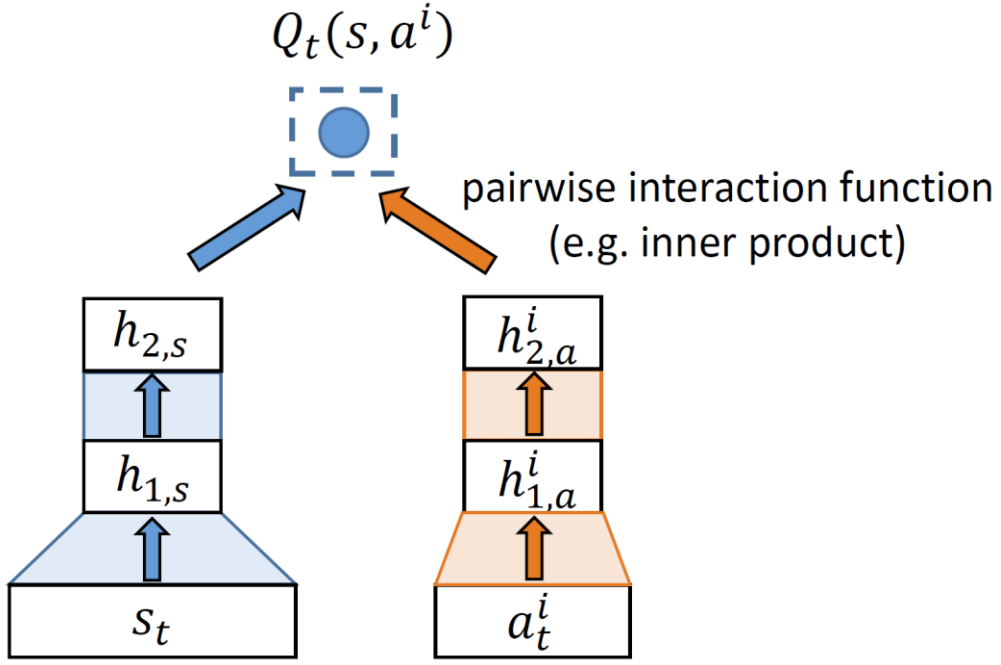
Deep Reinforcement Relevance Network

- **Deep Reinforcement Relevance Network** (DRRN)
  - Separate state and action embeddings
  - Interaction at the embedding space
  -

[He, Chen, He, Gao, Li, Deng, Ostendorf, ACL2016]

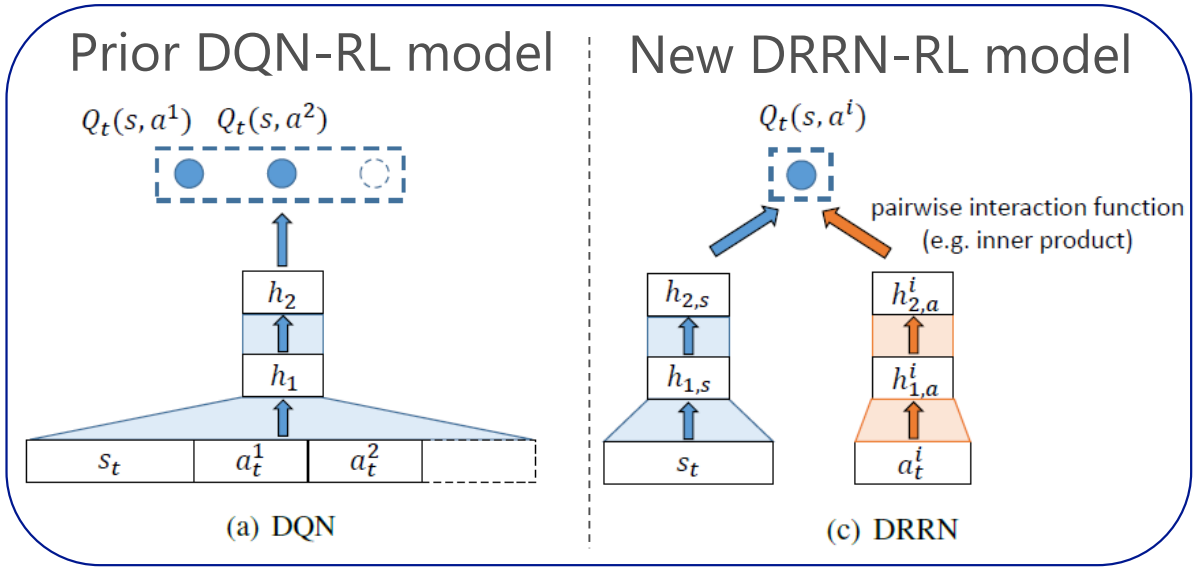
$$Q(s, a^i; \Theta) = g\left(h_{L,s},\ h^i_{L,a}\right)$$

Motivation:
- Language is different in these two contexts.
- Text similarity does NOT always lead to the best action.



Reflection: DRRN

- Prior DQN work (e.g., Atari game, AlphaGo): state space unbounded, **action space bounded**.
- In NLP tasks, usually **the action space is unbounded** since it is characterized by natural language, which is discrete and nearly unconstrained.
- New DRRN: (Deep Reinforcement Relevance Network)
  - Project both the state and the action into a continuous space
  - Q-function is an relevance function of the state vector and the action vector



[He, Chen, He, Gao, Li, Deng, Ostendorf, "Deep Reinforcement Learning with a Natural Language Action Space," ACL2016]

Experiments: Tasks

- Two text games

Stats	"Saving John"	"Machine of Death"
Text game type	Choice-based	Choice-based & Hypertext-based
Vocab size	1762	2258
Action vocab size	171	419
Avg. words/description	76.67	67.80
State transitions	Deterministic	Stochastic
# of states (underlying)	≥ 70	≥ 200
(Avg., max) steps/episode	14, ≥ 38	83, ≥ 500

- Hand annotate rewards for distinct endings
  - Simulators available at: https://github.com/jvking/text-games

Experiments

- Tasks: Text Games/Interactive Fictions
  - Task 1: "Saving John"

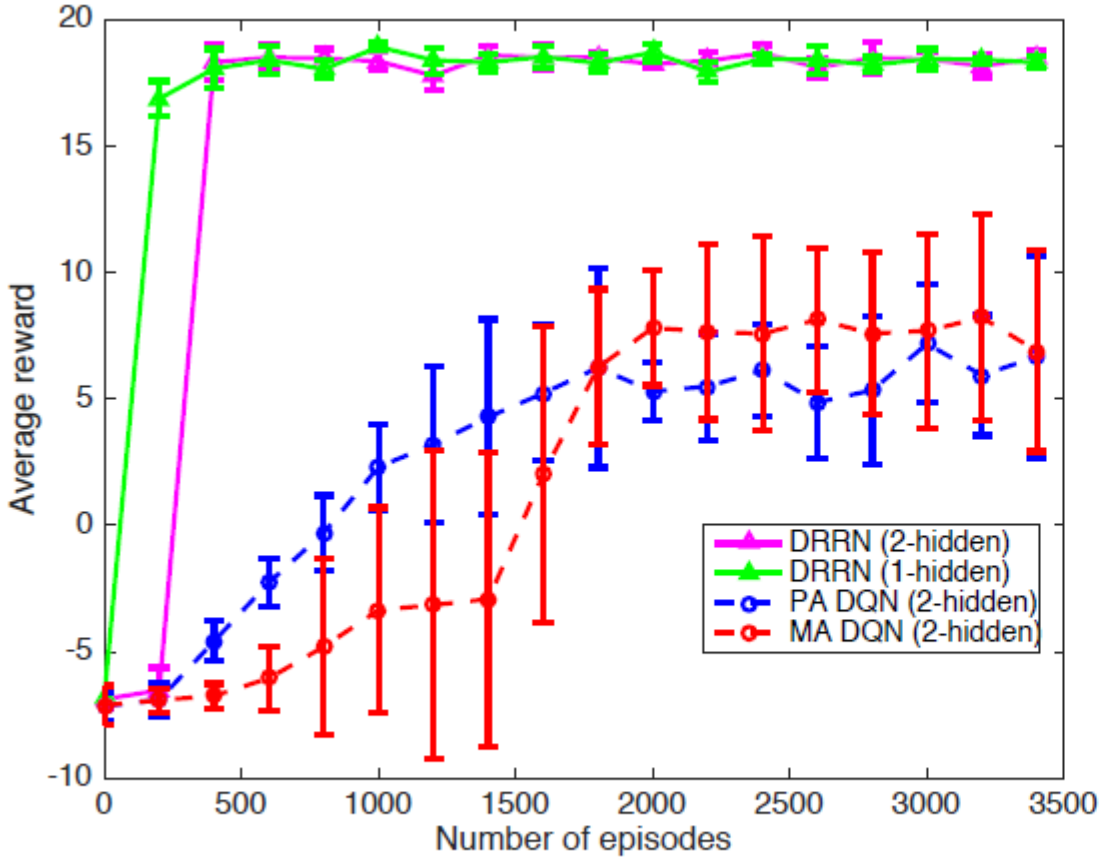
Reward	Endings (partially shown)
-20	Suspicion fills my heart and I scream. Is she trying to kill me? I don't trust her one bit...
-10	Submerged under water once more, I lose all focus...
0	Even now, she's there for me. And I have done nothing for her...
10	Honest to God, I don't know what I see in her. Looking around, the situation's not so bad...
20	Suddenly I can see the sky... I focus on the most important thing - that I'm happy to be alive.

Experiments

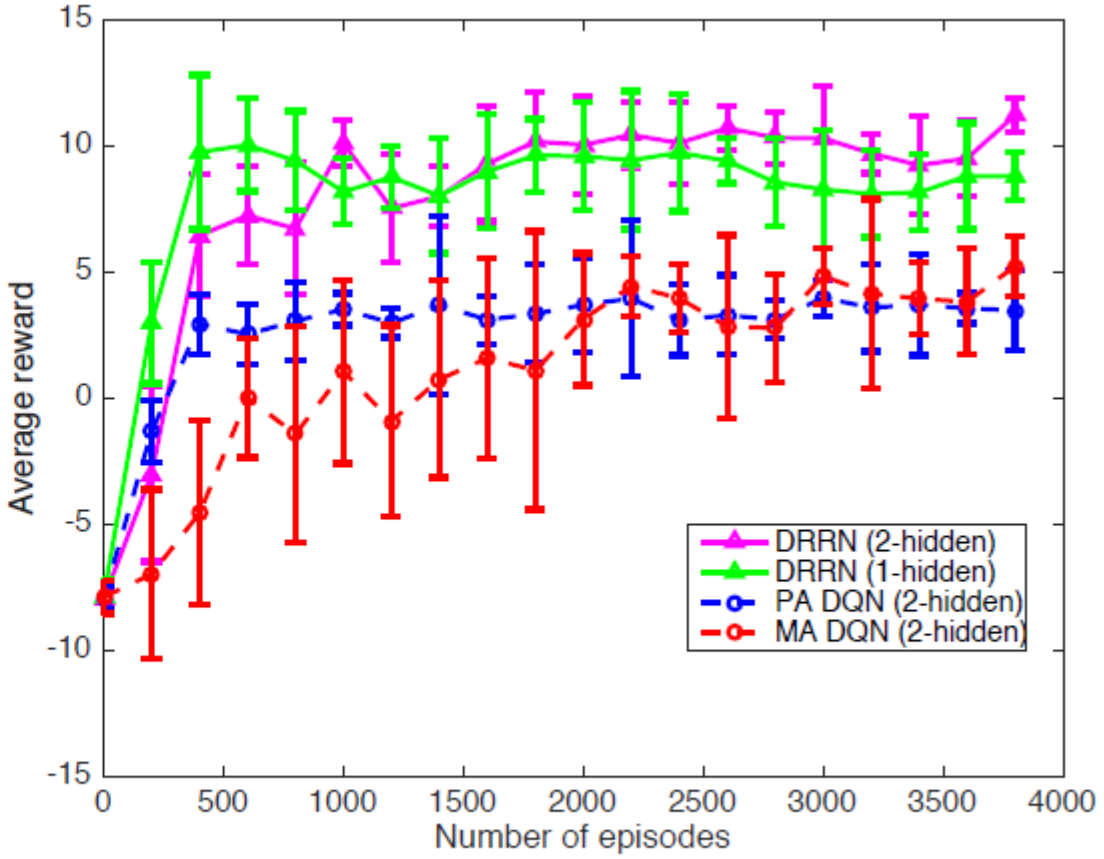
- Tasks: Text Games/Interactive Fictions
  - Task 2: "Machine of Death"

Reward	Endings (partially shown)
-20	You spend your last few moments on Earth lying there, shot through the heart, by the image of Jon Bon Jovi.
-20	you hear Bon Jovi say as the world fades around you.
-20	As the screams you hear around you slowly fade and your vision begins to blur, you look at the words which ended your life.
-10	You may be locked away for some time.
-10	Eventually you're escorted into the back of a police car as Rachel looks on in horror.
-10	Fate can wait.
-10	Sadly, you're so distracted with looking up the number that you don't notice the large truck speeding down the street.
-10	All these hiccups lead to one grand disaster.
10	Stay the hell away from me! She blurts as she disappears into the crowd emerging from the bar.
20	You can't help but smile.
20	Hope you have a good life.
20	Congratulations!
20	Rachel waves goodbye as you begin the long drive home. After a few minutes, you turn the radio on to break the silence.
30	After all, it's your life. It's now or never. You ain't gonna live forever. You just want to live while you're alive.

Learning curve: DRRN vs. DQN



(a) Game 1: "Saving John"

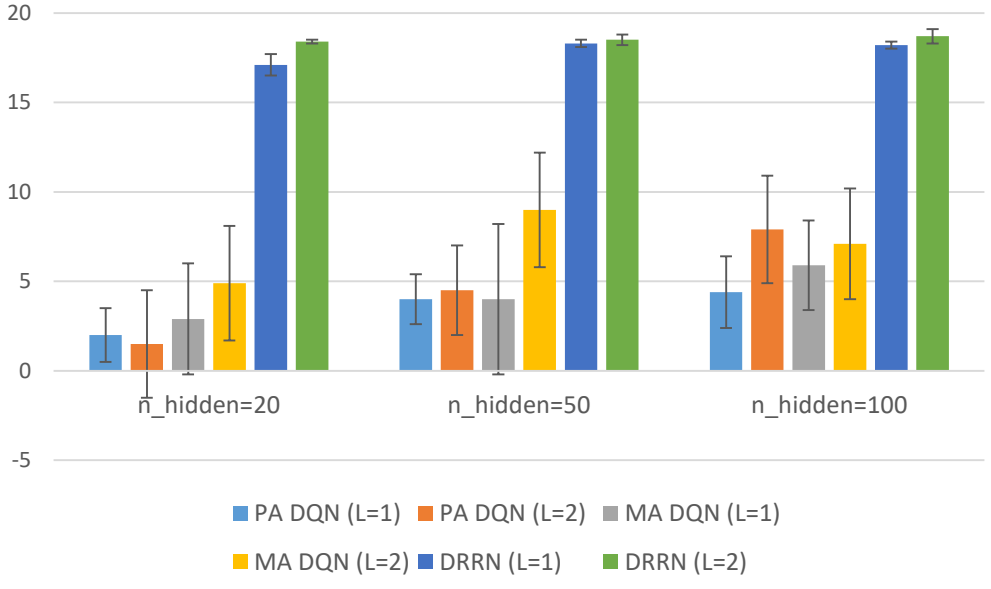


(b) Game 2: "Machine of Death"

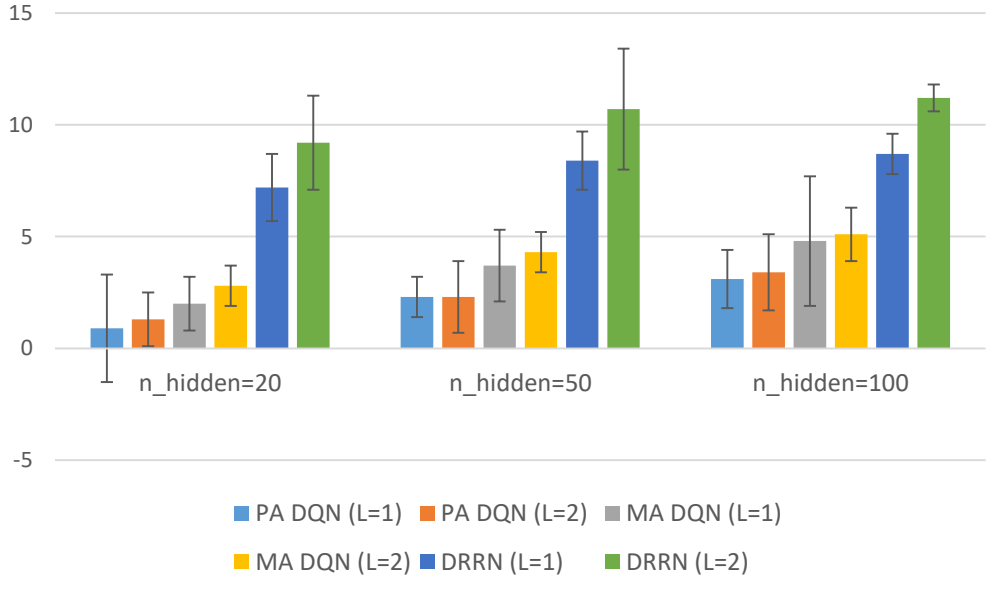
Tested on two text games

Experiments: Final Performance

Game 1: "Saving John"



Game 2: "Machine of Death"



The DRRN performs consistently better than all baselines, and often with a lower variance.

Big gain from having separate state & action embedding spaces (DQN vs. DRRN).

Visualization of the learned continuous space

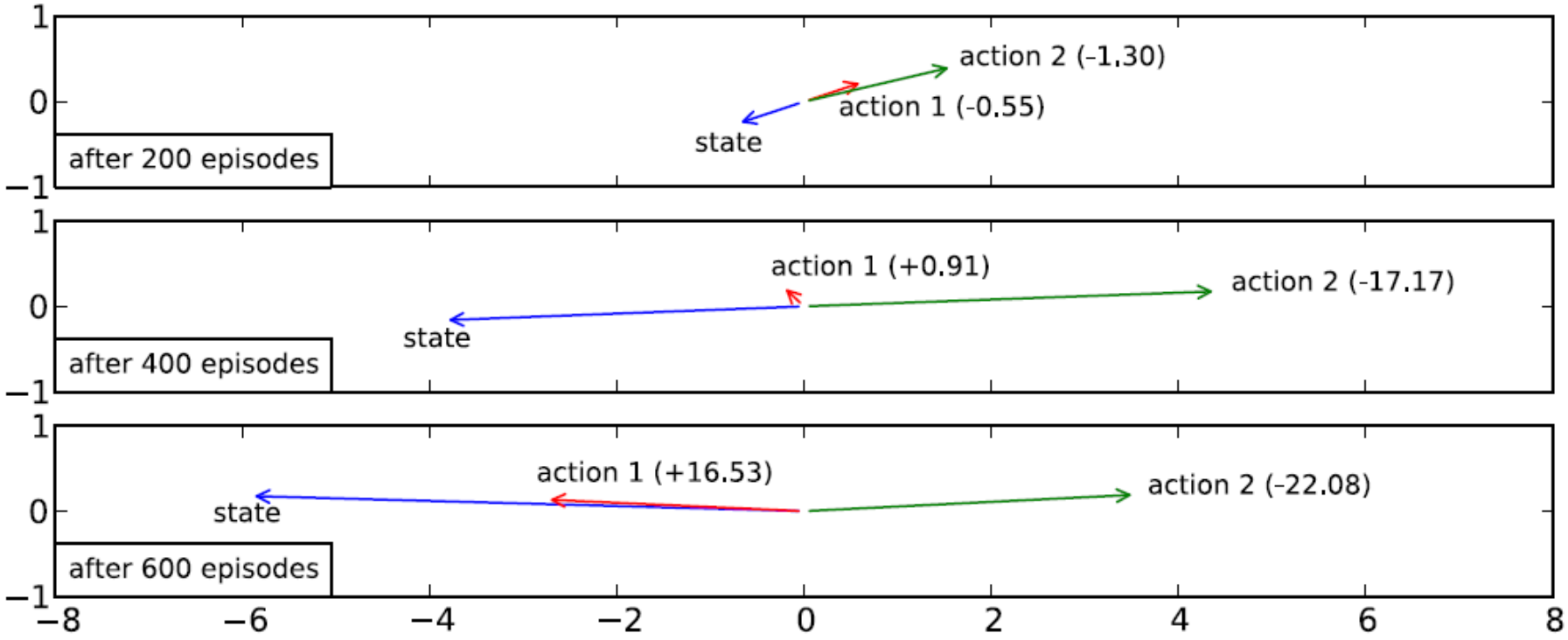
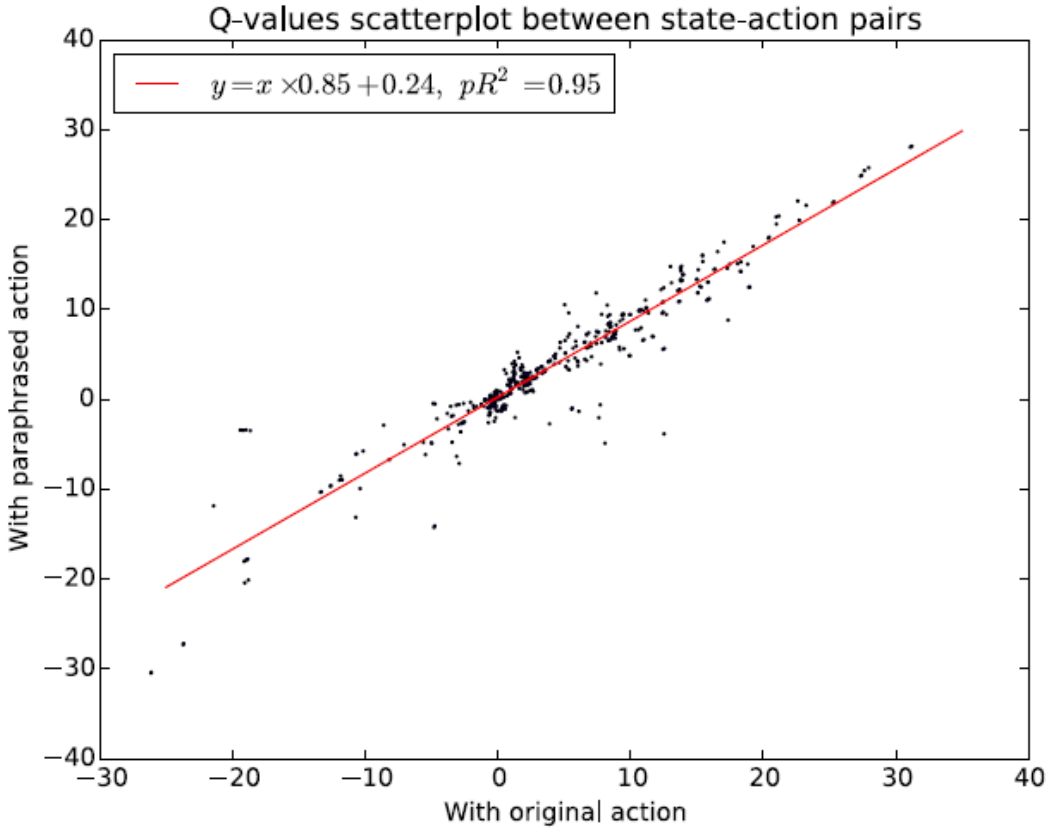
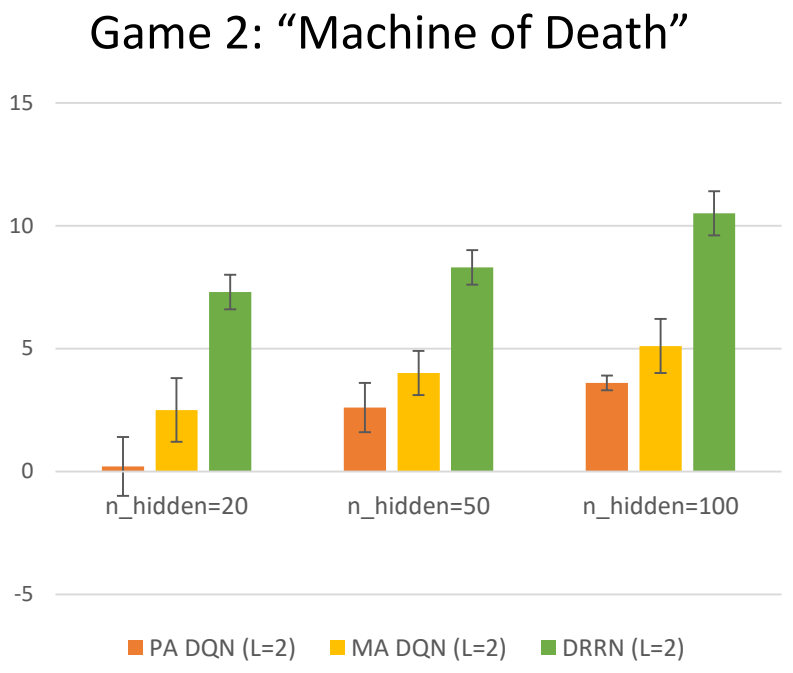


Figure 2: PCA projections of text embedding vectors for state and associated action vectors after 200, 400 and 600 training episodes. The state is "As you move forward, the people surrounding you suddenly look up with terror in their faces, and flee the street." Action 1 (good choice) is "Look up", and action 2 (poor choice) is "Ignore the alarm of others and continue moving forward."

Experiments: Generalization

- In the testing stage, use unseen paraphrased actions





Q-function example values after converged

	Text (with predicted Q-values)
State	As you move forward, the people surrounding you suddenly look up with terror in their faces, and flee the street.
Actions in the original game	Ignore the alarm of others and continue moving forward. (-21.5) Look up. (16.6)
Paraphrased actions (not original)	Disregard the caution of others and keep pushing ahead. (-11.9) Turn up and look. (17.5)
Fake actions (not original)	Stay there. (2.8) Stay calmly. (2.0) Screw it. I'm going carefully. (-17.4) Yell at everyone. (-13.5) Insert a coin. (-1.4) Throw a coin to the ground. (-3.6)

Note that, the DRRN generalizes to unseen actions well, e.g., for these "not original" actions, the model still gives a proper estimate of the Q-value.

Latest advances

- Our next EMNLP2016 paper extends the work to real-world large NLP data set, and targeting on combinatorial action spaces
  - Deep Reinforcement Learning with a Combinatorial Action Space for Predicting and Tracking Popular Discussion Threads
  - https://arxiv.org/abs/1606.03667

  The agent runs on real world ***Reddit*** dataset https://www.reddit.com/

  reads Reddit posts

  recommends most thread with most future potential popularity

Related Work

- Recently, significant progress has been made by combining deep learning with RL (Mnih et al. 2015, Silver et al. 2016)
- Narasimhan et al. (2015) introduced deep RL in parser-based text games
- Nogueira and Cho (2016) proposed a goal-driven web navigation task for language based sequential decision making study
- Narasimhan et al. (2016) applied RL for acquiring and incorporating external evidence to improve information extraction accuracy
- Very recently, in RL applying to human-computer dialogue system, efforts from various research labs: University of Cambridge, MSR, Stanford University, and Maluuba

Interim summary

Reinforcement learning for NLP tasks in a continuous space

- Project both states and actions (defined by *unbounded* NL) to a continuous semantic space using deep neural nets
- Compute the Q function in the continuous semantic space

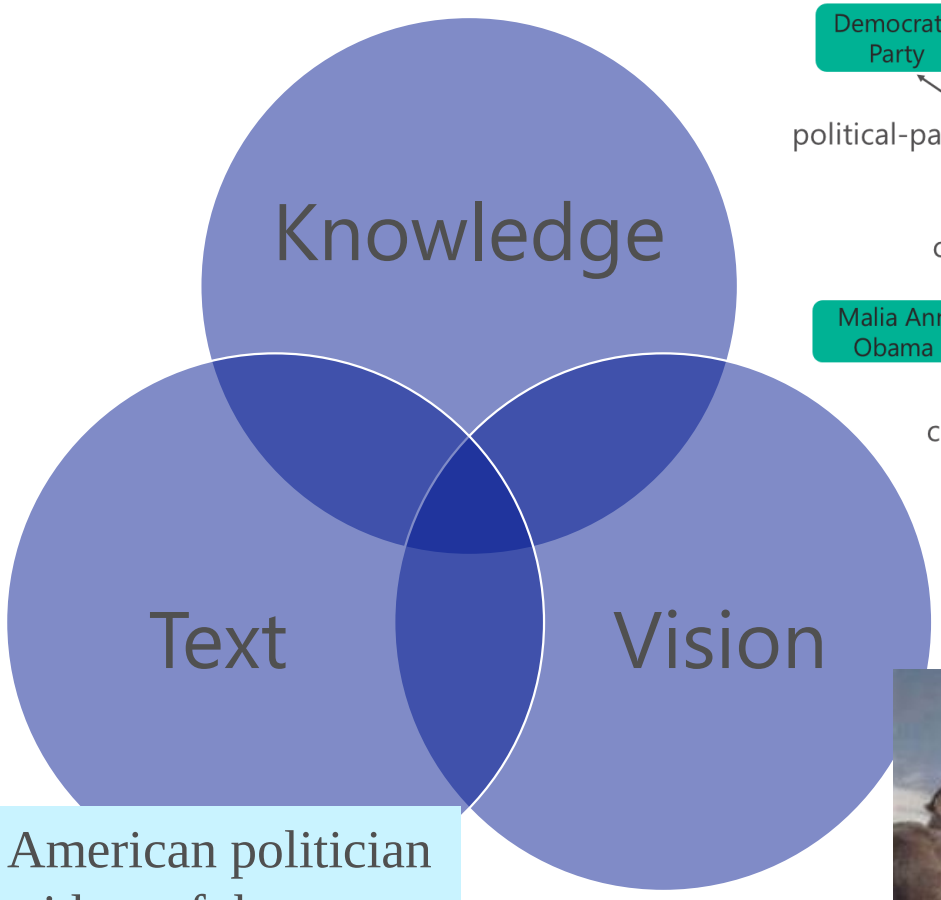
Part VI

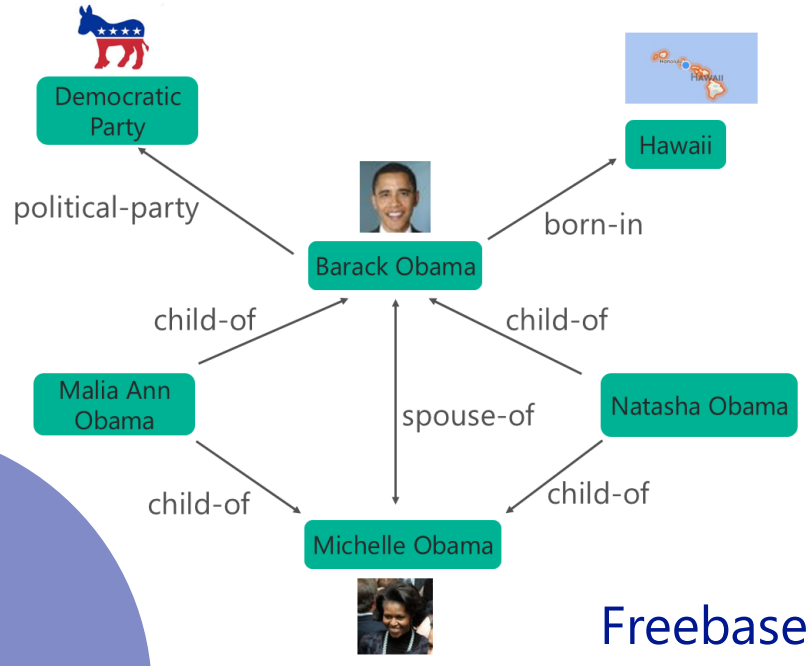
Image-language multimodal learning and inference

Image-language Joint learning and inference

- Image Captioning
- Visual Question answering

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)

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

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!

Image Captioning: Understanding complex scenes

-- Let's do a Turing Test!

a bicycle is parked next to a river

a bike sits parked next to a body of water

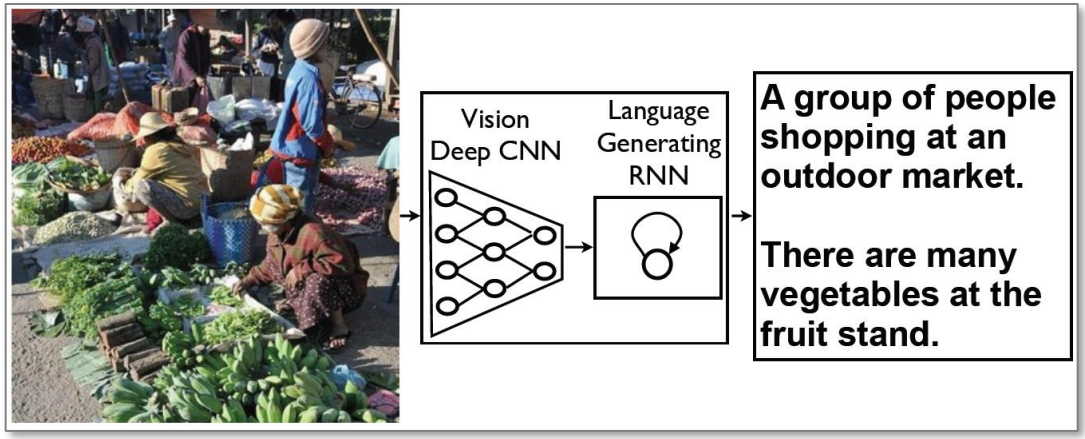
?

Two major paradigms

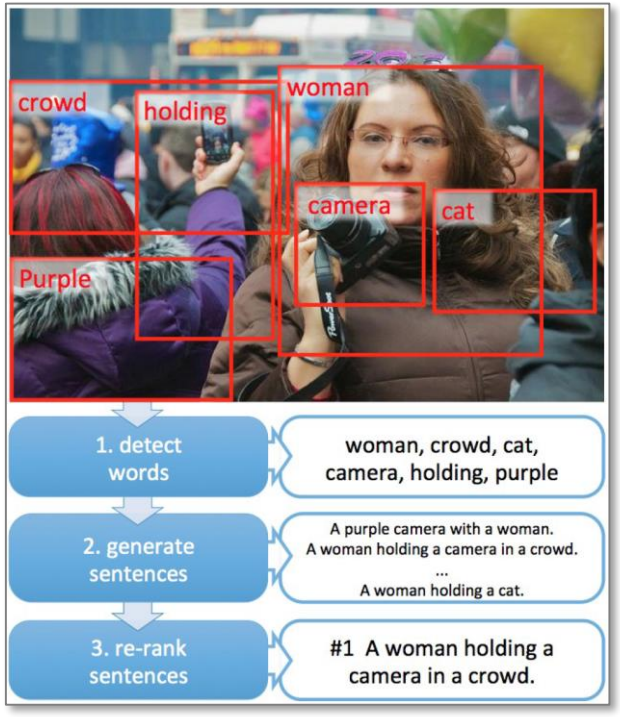
Two entries tied at the 1st place at COCO 2015 Caption Challenge

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

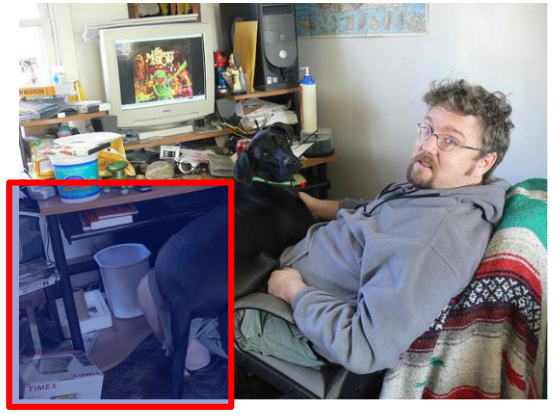
$$p_i^w = 1 - \prod_{j \in r_i} (1 - \overbrace{\sigma(f_{ij} \cdot v_w)}^{p(w \text{ in } r_j \text{ of image } i)})$$

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**.

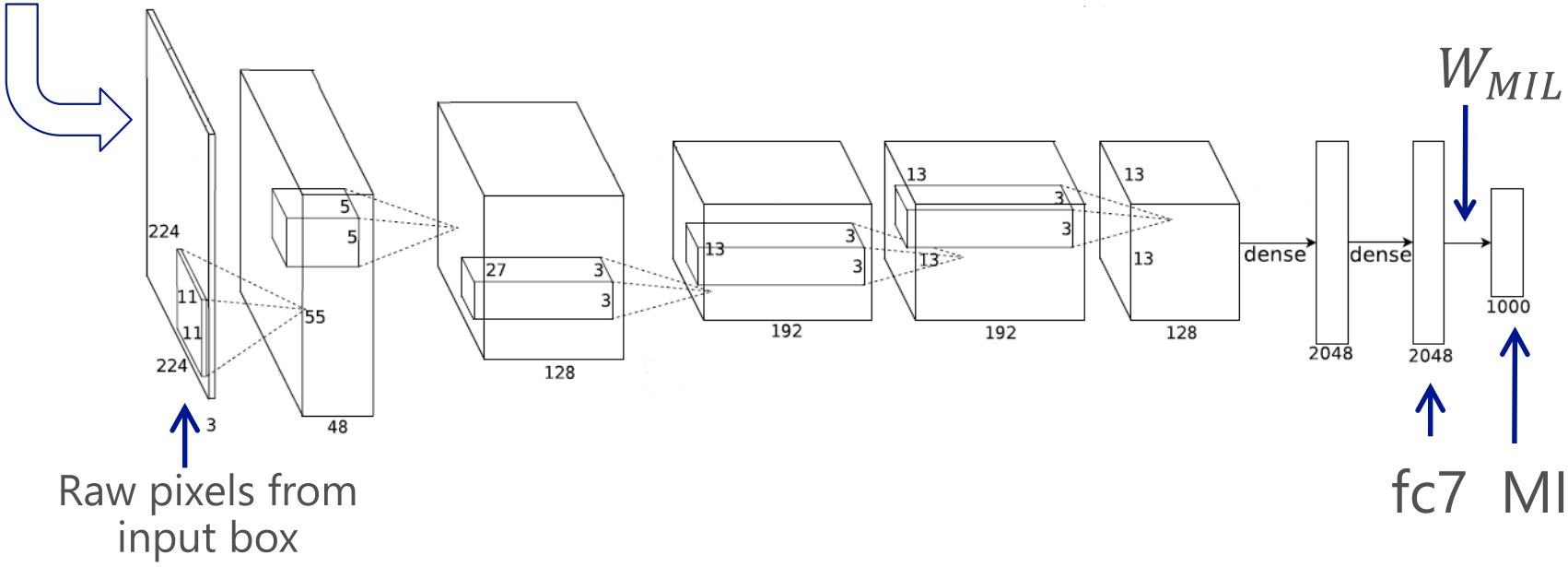
$$h(x, y) = \sum_{r_i, s.t., (x,y) \in r_i} \sigma(f_{ij} \cdot v_{sitting})$$

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.

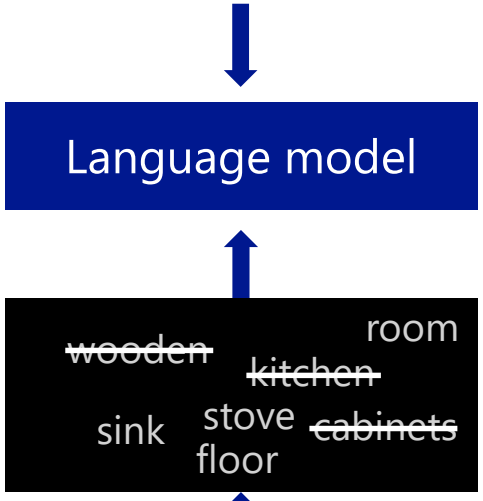
$$\Pr(w_l = \bar{w}_l | \bar{w}_{l-1}, \cdots, \bar{w}_1, <s>, \tilde{\mathcal{V}}_{l-1}) = \frac{\exp\left[\sum_{k=1}^{K} \lambda_k f_k(\bar{w}_l, \bar{w}_{l-1}, \cdots, \bar{w}_1, <s>, \tilde{\mathcal{V}}_{l-1})\right]}{\sum_{v \in \mathcal{V} \cup </s>} \exp\left[\sum_{k=1}^{K} \lambda_k f_k(v, \bar{w}_{l-1},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

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



cabinets

MaxEnt LM

$p(cabinets|with\ wooden)$

a kitchen with wooden cabinets

Image

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]

Multimodal DSSM

- Project sentence and image into a comparable semantic vector space
- Whole sentence language model

Q = image, D = caption, R = relevance

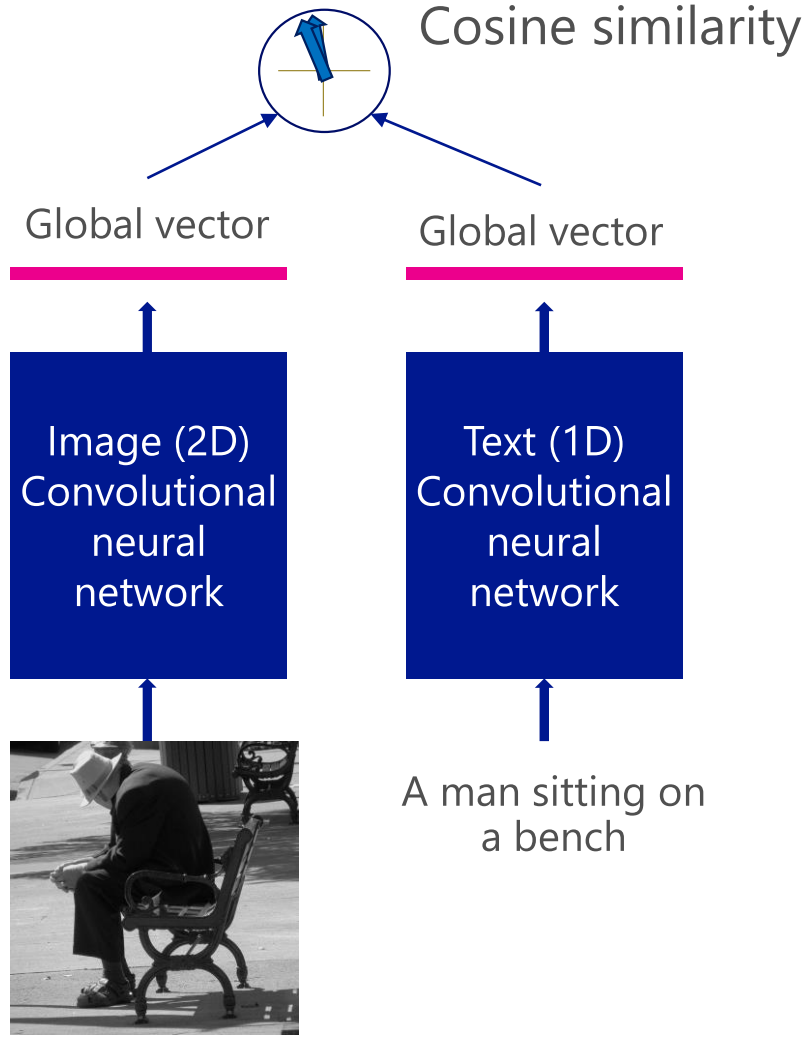
Relevance: $R(Q,D) = \text{cosine}(y_Q, y_D) = \frac{{y_Q}^T y_D}{\|y_Q\|\|y_D\|}$

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

Candidate captions ↗ ↖ Smoothing factor

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

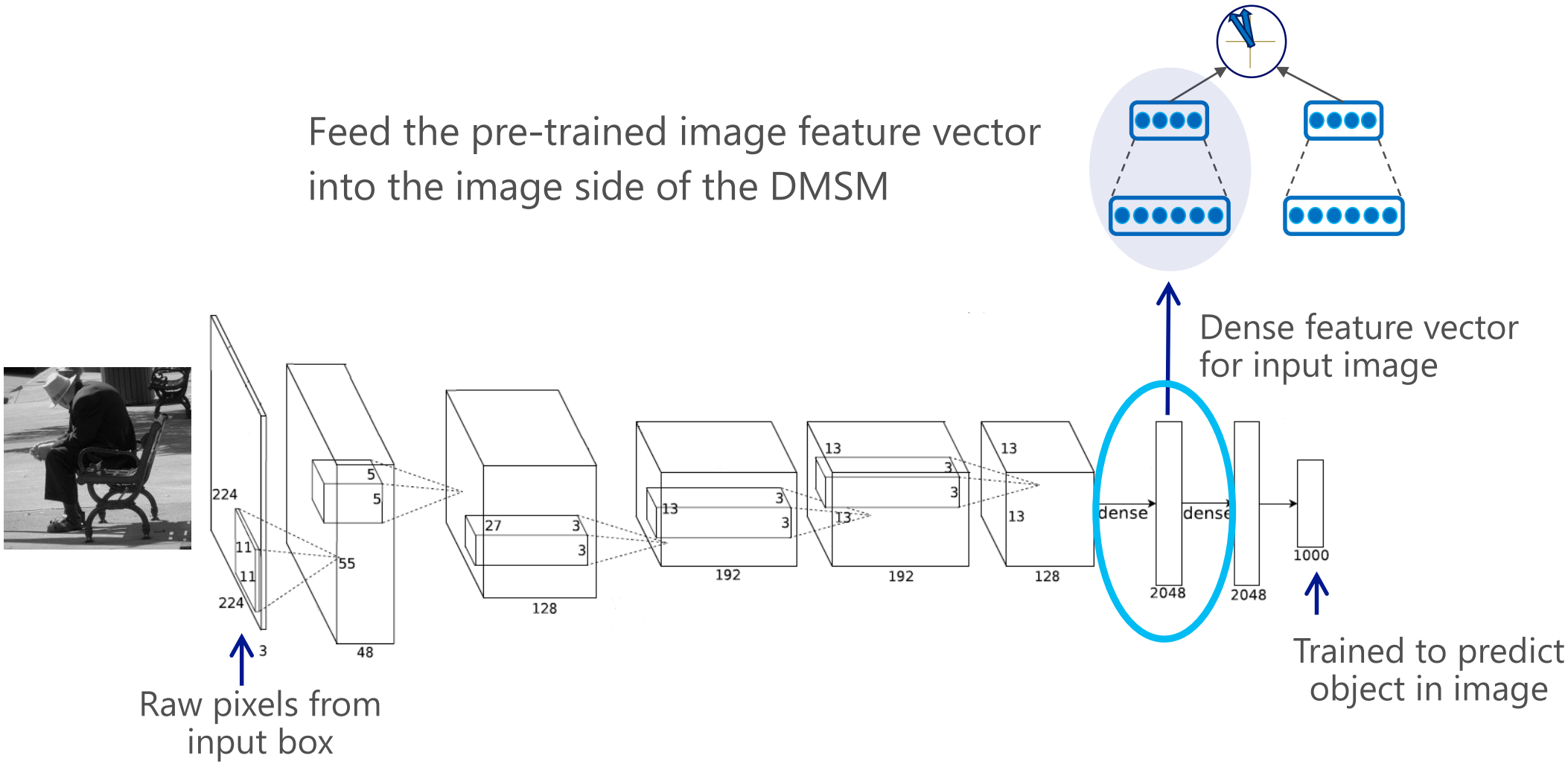
↖ Correct caption



Serves as a semantic matching checker.

The convolutional network at the image side

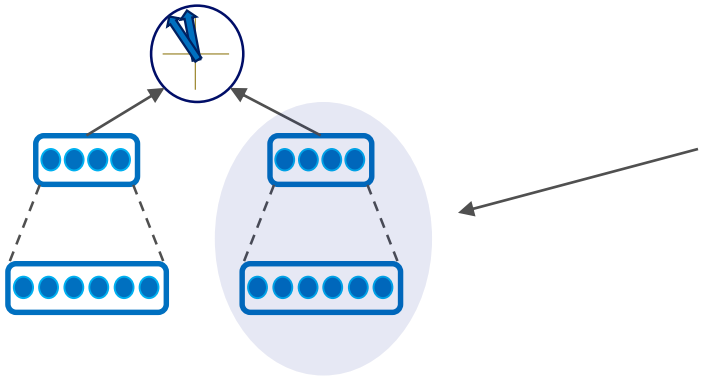
Feed the pre-trained image feature vector into the image side of the DMSM



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

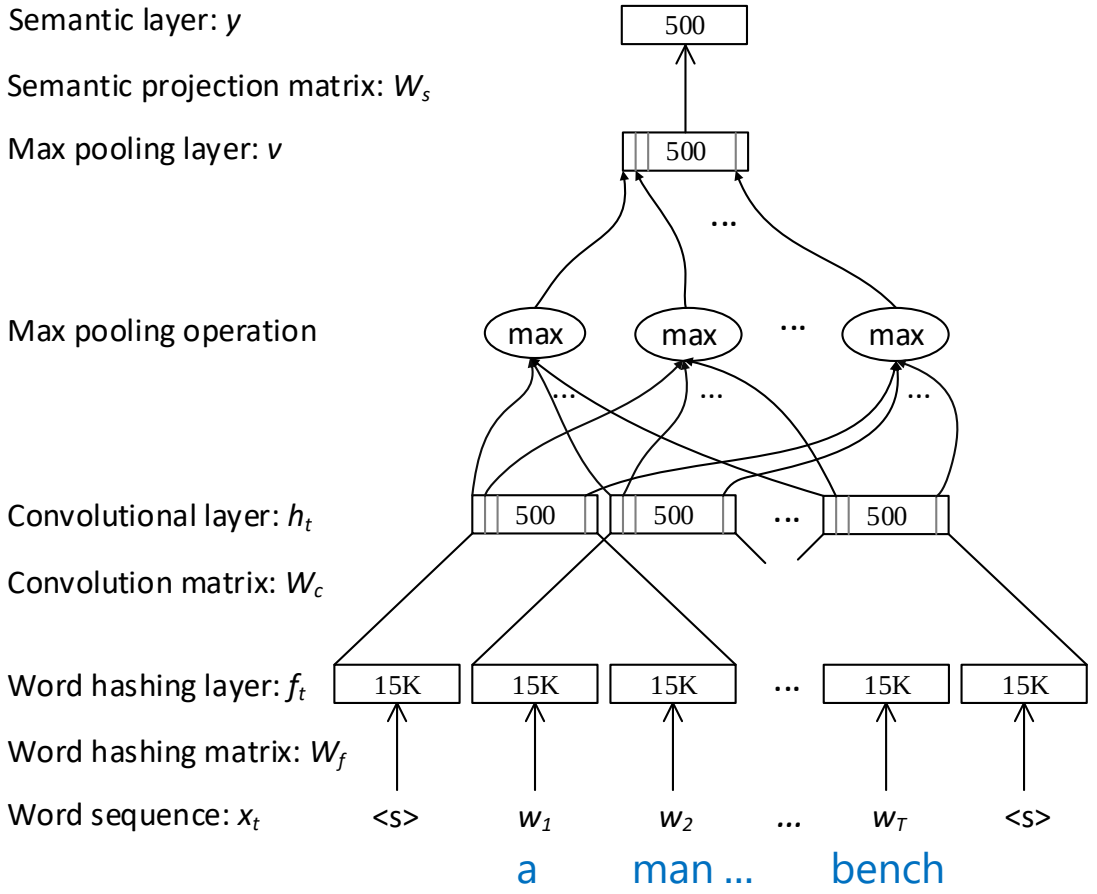


Figure Credit: [Shen, He, Gao, Deng, Mesnil, WWW, April 2014]

State of the art

Human judgment is the ultimate metric

Turing Test Results at the MS COCO Image Captioning Challenge 2015

MSR won the 1st prize!

	Official Rank	% of captions that pass the Turing Test	% of captions that are better or equal to human's
MSR	1st	32.2%	26.8%
Google	1st	31.7%	27.3%
MSR Captivator	3rd	30.1%	25.0%
Montreal/Toronto	3rd	27.2%	26.2%
Berkeley LRCN	5th	26.8%	24.6%
Other groups: Baidu/UCLA, Stanford, Tsinghua, etc.			
Nearest neighbor	--	**25.5%**	**21.6%**
Human	--	**67.5%**	**63.8%**

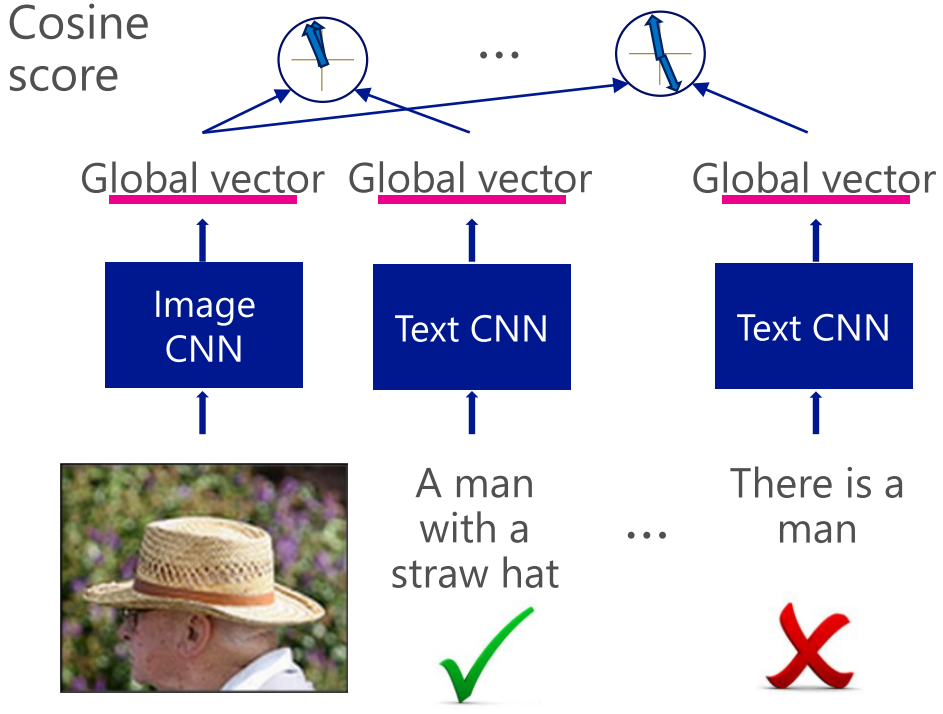
Retrieval-based method

human generated caption

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



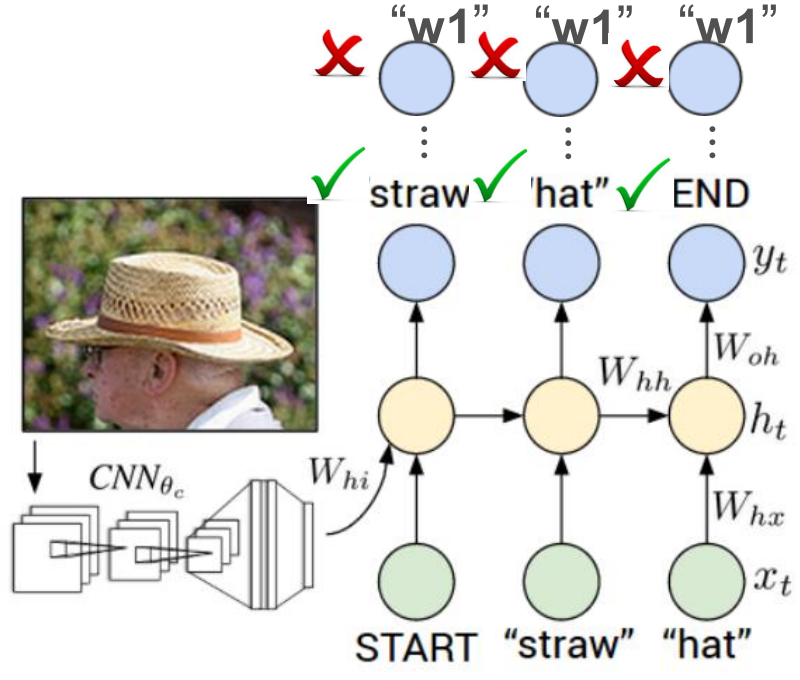


Image Credit: Karpathy and Fei-Fei 2015

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

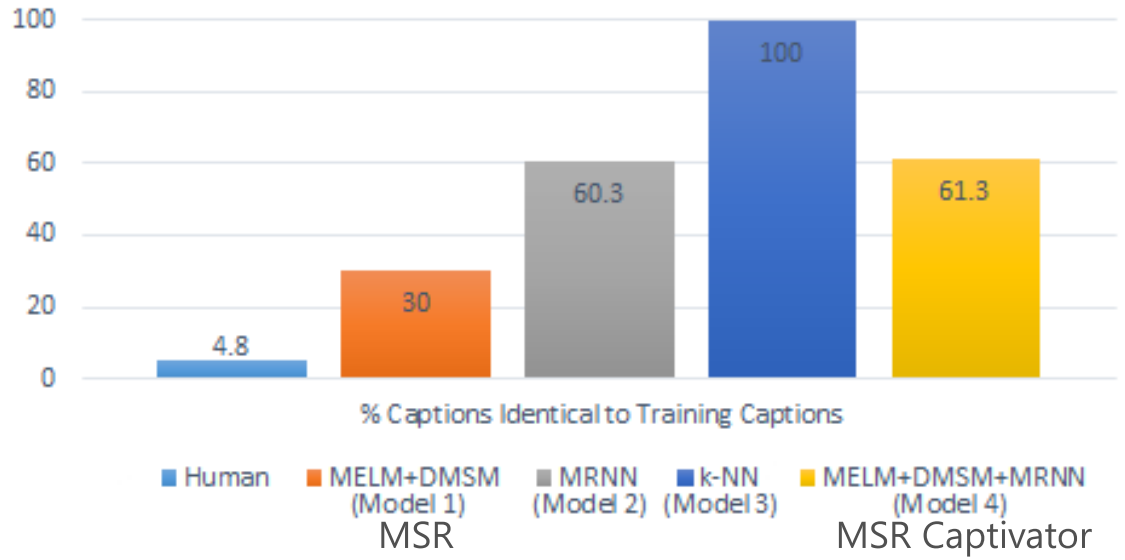
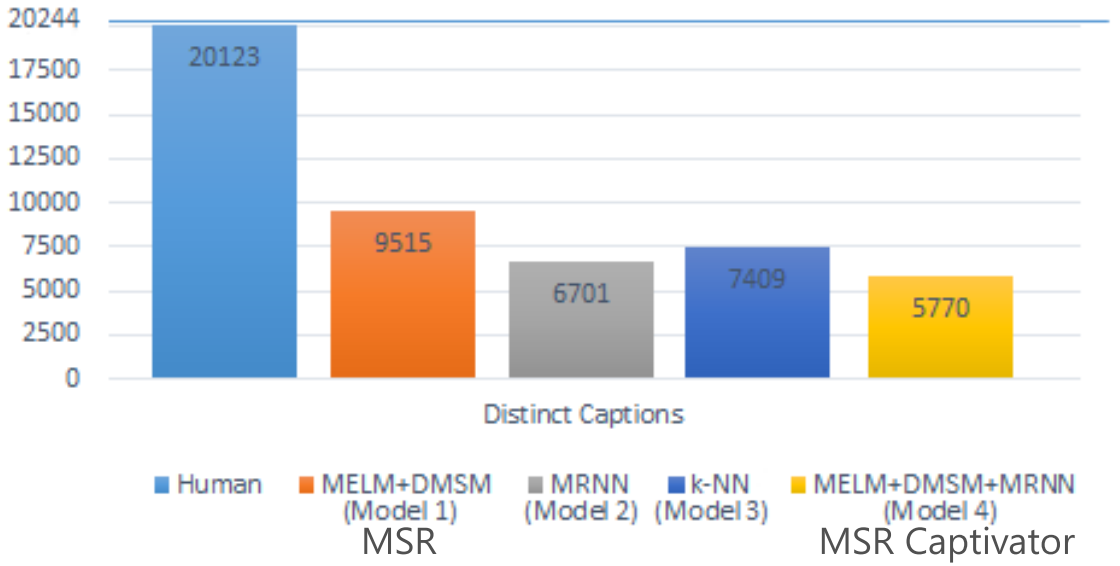


Table 2: Number *Distinct* Captions in Testval (out of 20,244 instances)



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

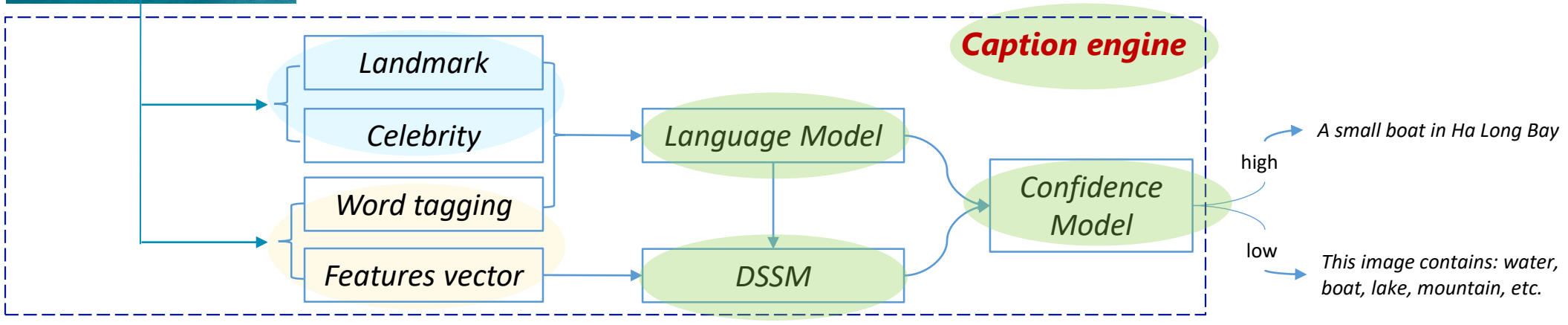
MRNN: "A close up of a plate of food" *(more generic)*

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

New architecture and system development



Codename: Cape Cod

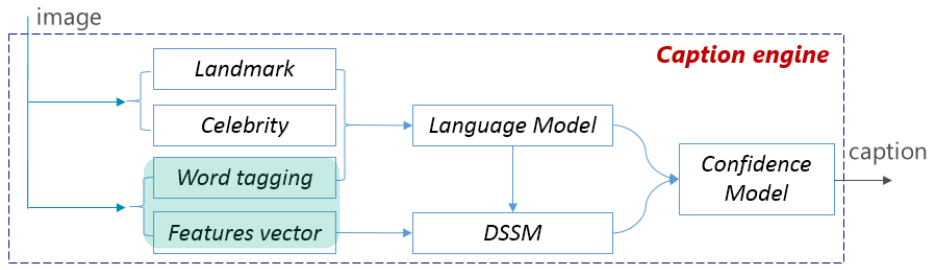
[Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, Chris Sienkiewicz, "*Rich Image Captioning in the Wild*," Deep Vision, CVPR, 2016]

An example - image captioning

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

(note the system missed Marian Robinson)

Deep ResNet for visual concepts detection

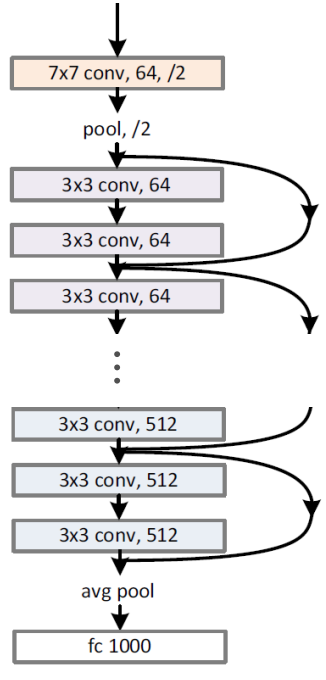


[He, Zhang, Ren, Sun, 2015]

ImageNet 2015 Winner !

ResNet

- Treat as multiclass problem
- Sigmoid output
- No softmax normalization
- Trained on multiple GPUs

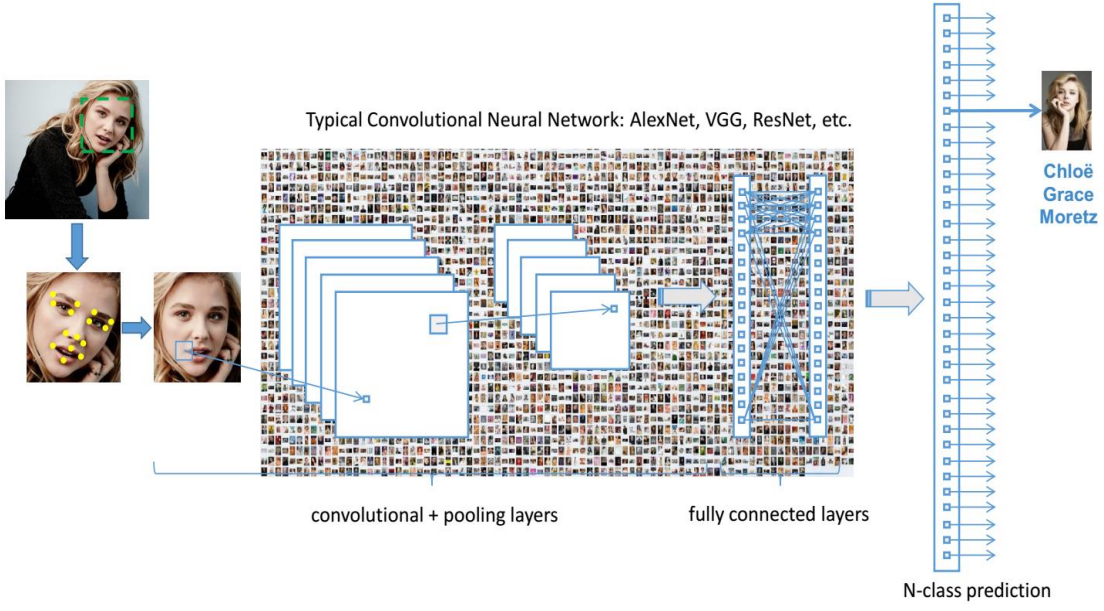


man, tennis, court, holding, shirt, yellow, racquet, …

also extract the 2nd last layer as feature representation.

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."

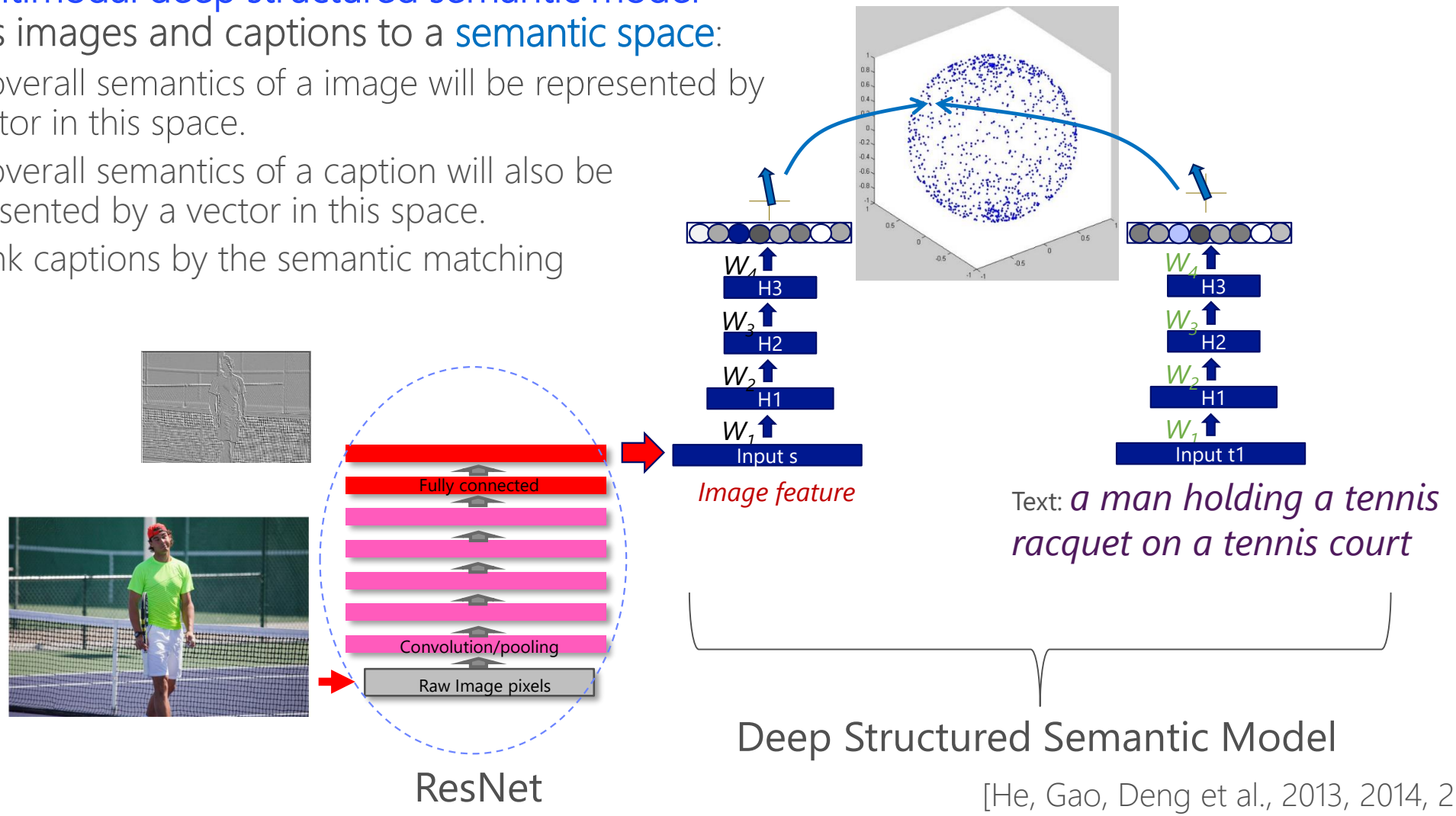
- Integrating entities (celebrities, landmarks, etc.) makes captions much richer.

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

DSSM: Bridge the gap between image and language!

The multimodal deep structured semantic model projects images and captions to a 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.
- Rerank captions by the semantic matching

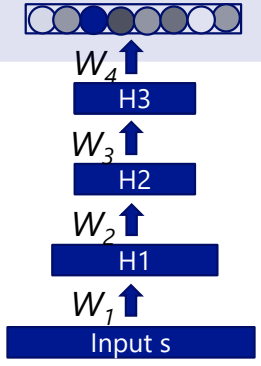


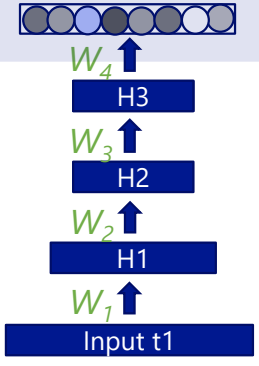
[He, Gao, Deng et al., 2013, 2014, 2015]

Describe with uncertainty

Confidence score [0,1] $s = \frac{e^{W \cdot f}}{1 + e^{W \cdot f}}$

1. LM score
2. Length
3. LM score per word
4. Vision tag covered
5. DSSM score





Image

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

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

Service APIs shipped in March 2016

Computer Vision API

Extract rich information from images to categorize and process visual data—and protect your users from unwanted content.

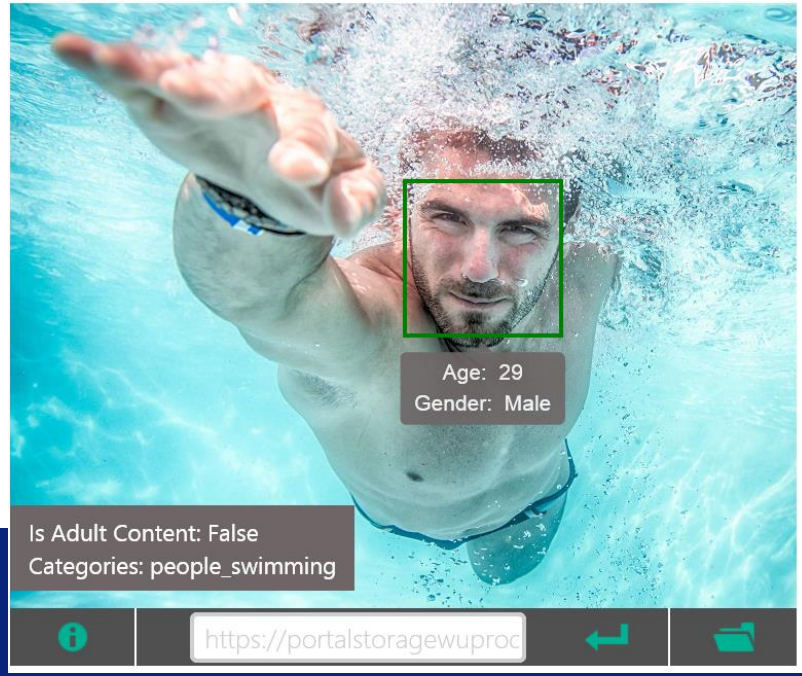
Microsoft Cognitive Services

Analyze an image

This feature returns information about visual content found in an image. Use tagging, descriptions and domain-specific models to identify content and label it with confidence.

https://www.microsoft.com/cognitive-services/en-us/computer-vision-api

"a man swimming in a pool of water"



Features:

Feature Name	Value
Description	{ "type": 0, "captions": [{ "text": "a man swimming in a pool of water", "confidence": 0.7850108124440484 }] }
Tags	[{ "name": "water", "confidence": 0.9996442794799805 }, { "name": "sport", "confidence": 0.9504992365837097 }, { "name": "swimming", "confidence": 0.9062818288803101, "hint": "sport" }, { "name": "pool", "confidence": 0.8787588477134705 }, { "name": "water sport", "confidence": 0.631849467754364, "hint": "sport" }]
Image Format	jpg

Positive feedback since announced

- CaptionBot & API highlighted at Satya's keynote at //Build2016

"Microsoft's newest bot offered a spot-on caption to this photo of Satya Nadella" – *Business Insider (Julie Bort)*

"Microsoft created Captionbot.ai, which is a tremendously addictive (and science-fiction-grade awesome)." -- *TechCrunch*

- Microsoft researchers tie for best image captioning technology *TechNet*
- Microsoft reveals 'CaptionBot' you can try out online - *Dailymail*
- Microsoft's Spooky New Bot Can Automatically Caption Your Photos - *Forbes*
- Microsoft's latest AI party trick is a CaptionBot for photos – *PCWorld*
- …

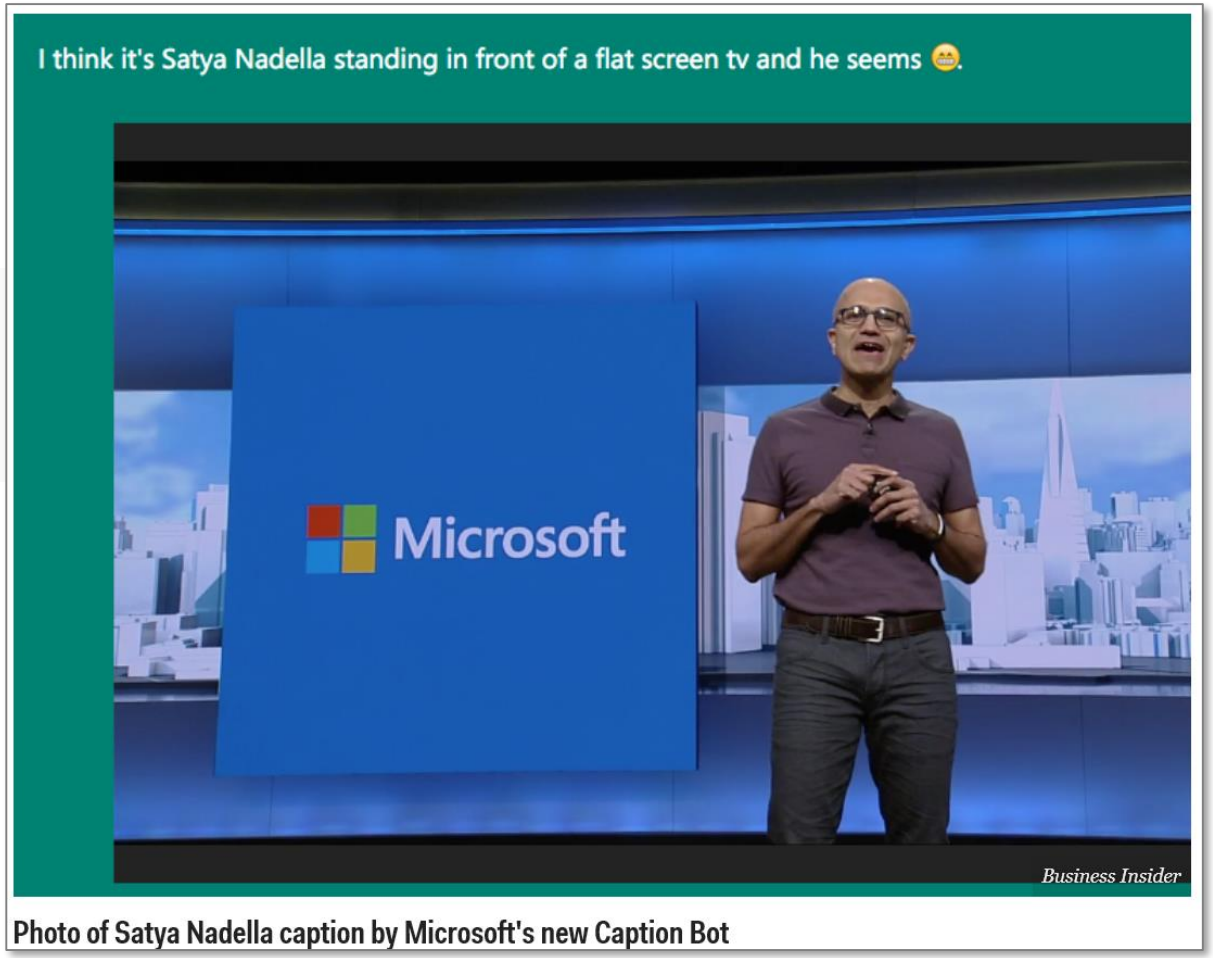
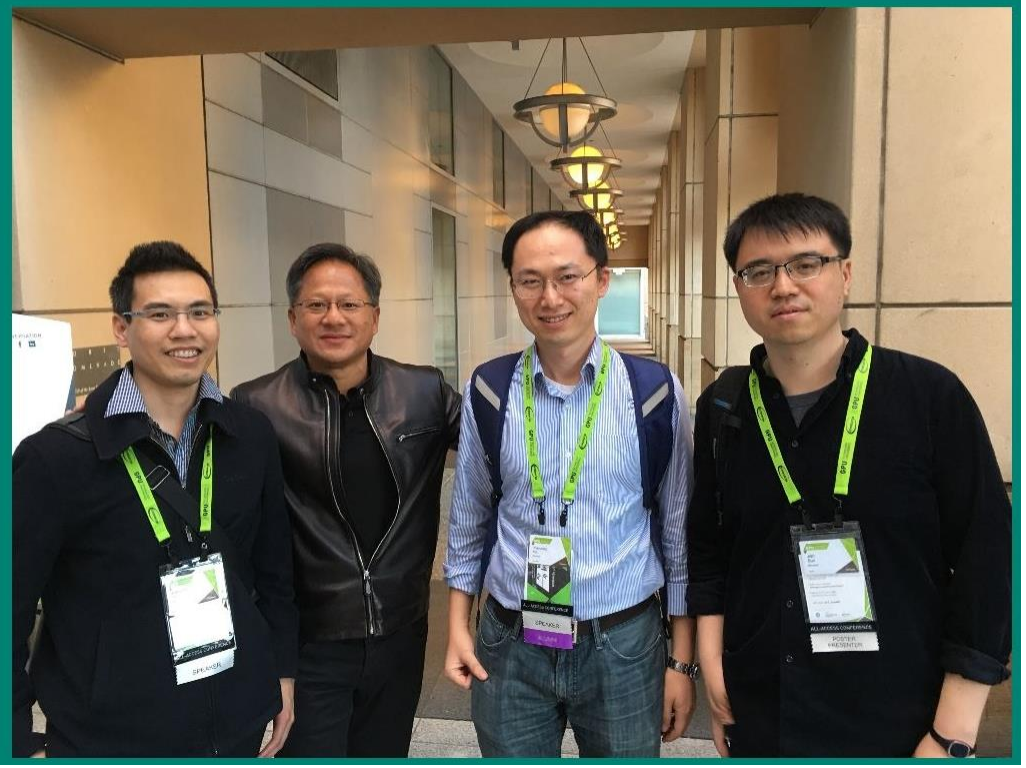


Photo of Satya Nadella caption by Microsoft's new Caption Bot

More examples from GTC2016

I think it's a group of people standing in front of a mountain.

I think it's Jen-Hsun Huang et al. that are posing for a picture and they seem 😐😁😁😁.



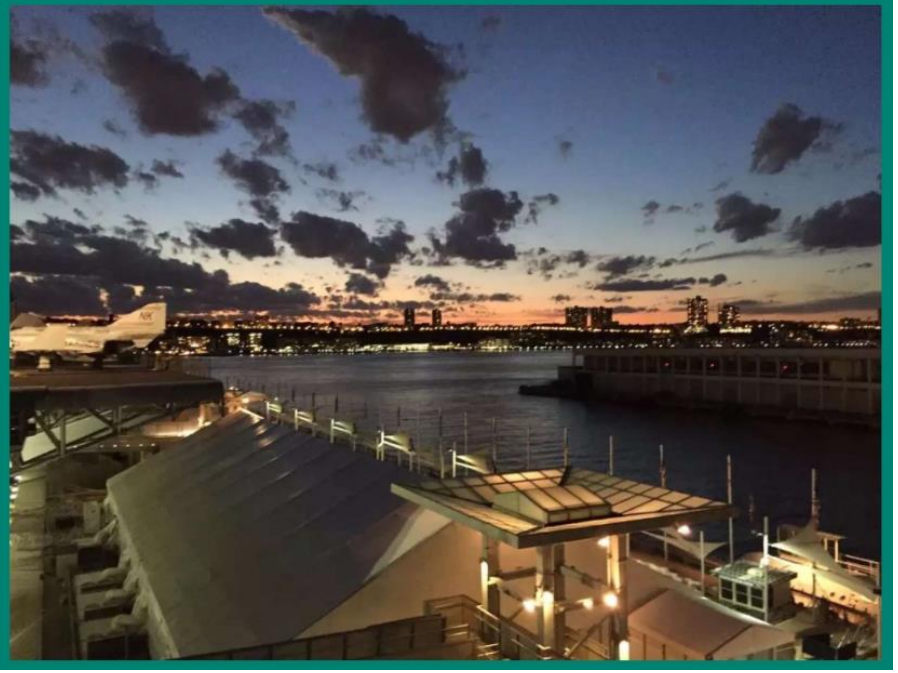
when Jen-Hsun Huang (NVidia CEO) was giving a keynote at GTC16 showing off a GPU-powered VR visiting of Mt. Everest -- here is what our CaptionBot has to say.

Public App: CaptionBot

http://CaptionBot.ai

works with any phone/browser

I think it's a boat that is lit up at night in a city.



AT&T 12:47 AM 59%

captionbot.ai

Microsoft

CaptionBot

I think it's a crowd of people walking down a street next to tall buildings and they seem 😐😐😐.

More examples from CaptionBot

CaptionBot

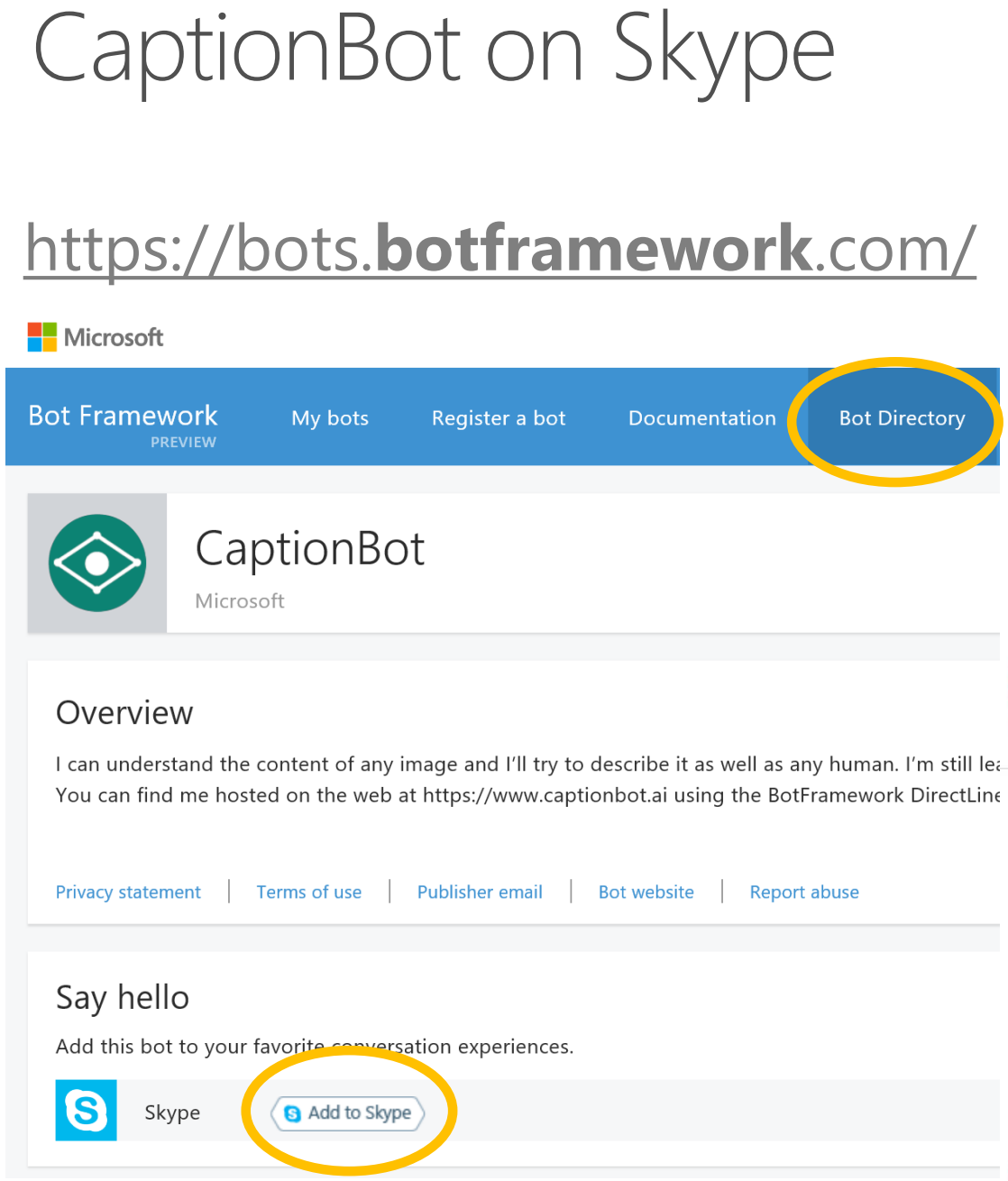
I am not really confident, but I think it's Leonardo da Vinci sitting in front of a mirror and she seems 😐.

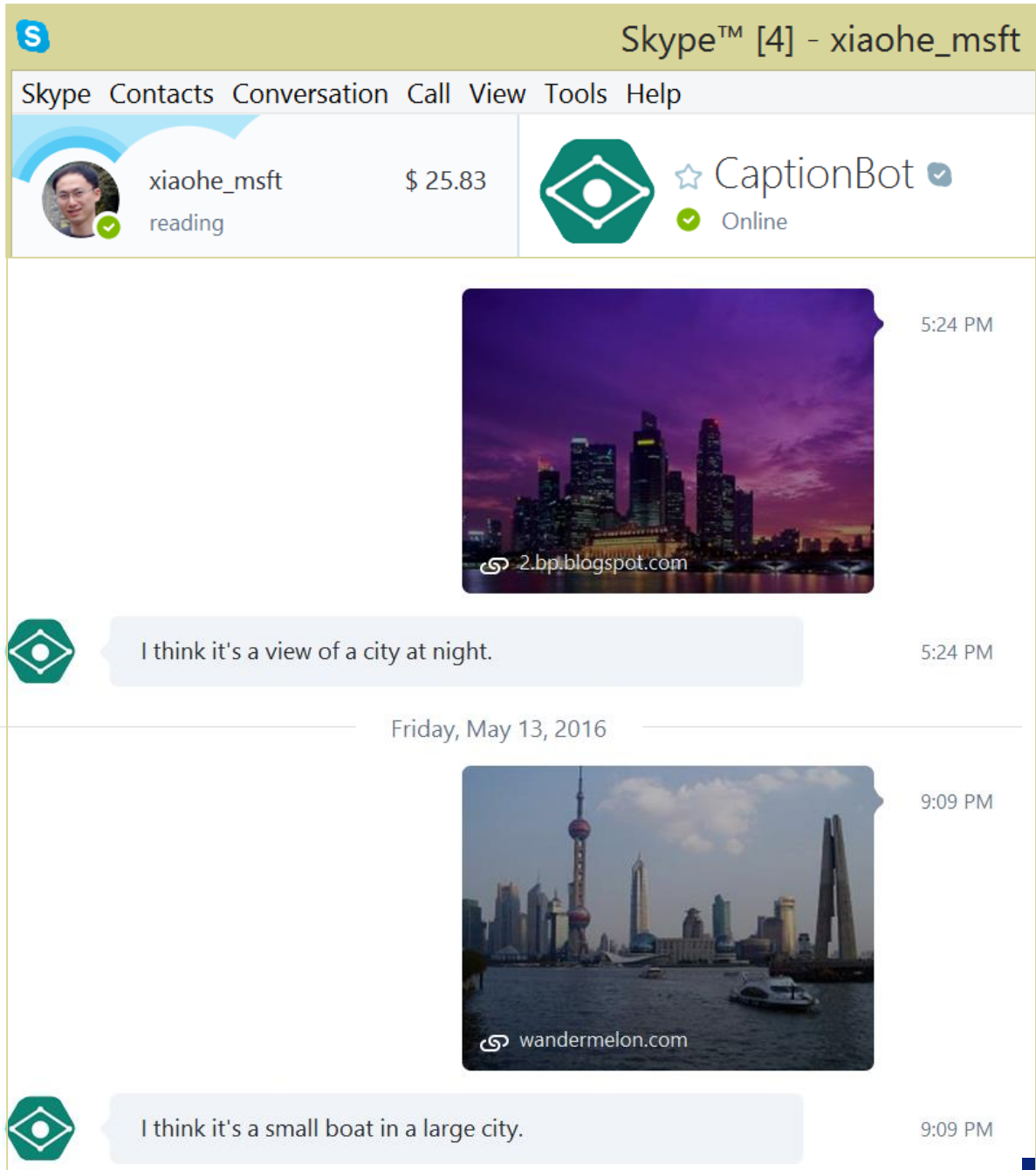
y street filled with

I think it's a man standing in front of a building and he seems 😁.

CaptionBot on Skype

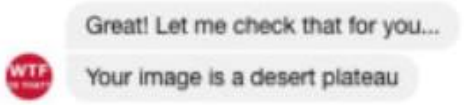
https://bots.**botframework**.com/





World-best vision ability in everyone's hands

By calling our API, a young guy, Ming Cheuk, at New Zealand made a computer vision bot WTF Is That on Facebook Messenger that's gone viral.

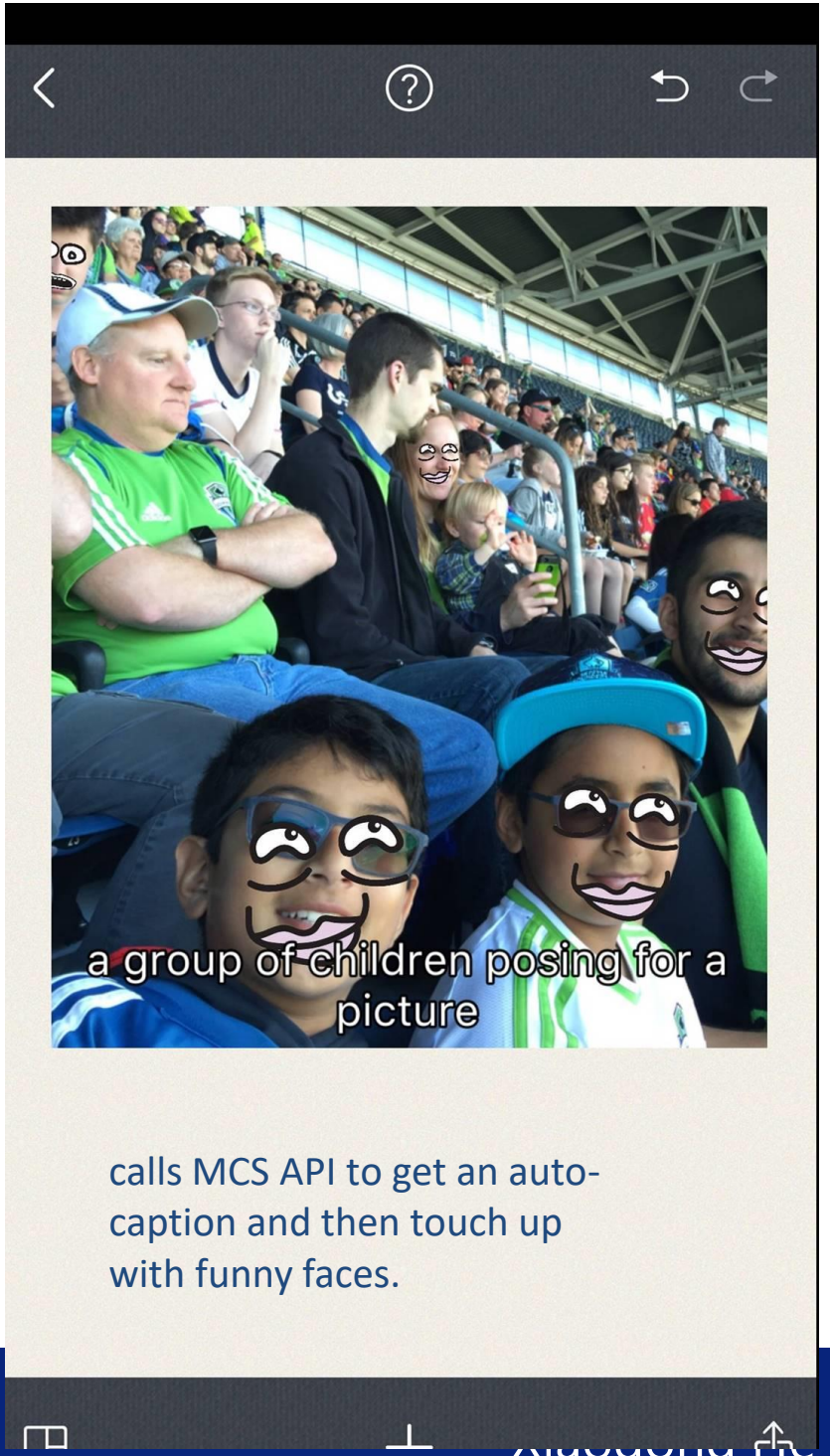


When a user uploads a photo, the bot uses Microsoft Cognitive Services' API to analyze the image and offer a response. Cheuk says Microsoft's tool provided the greatest scaleability, but he's testing services like Google Cloud Vision API, CloudSight, and Clarifai.

http://www.wired.com/2016/05/wtf-is-that-computer-vision-bot/

Enable 3rd party apps

PicCollage (15M users and growing)



calls MCS API to get an auto-caption and then touch up with funny faces.

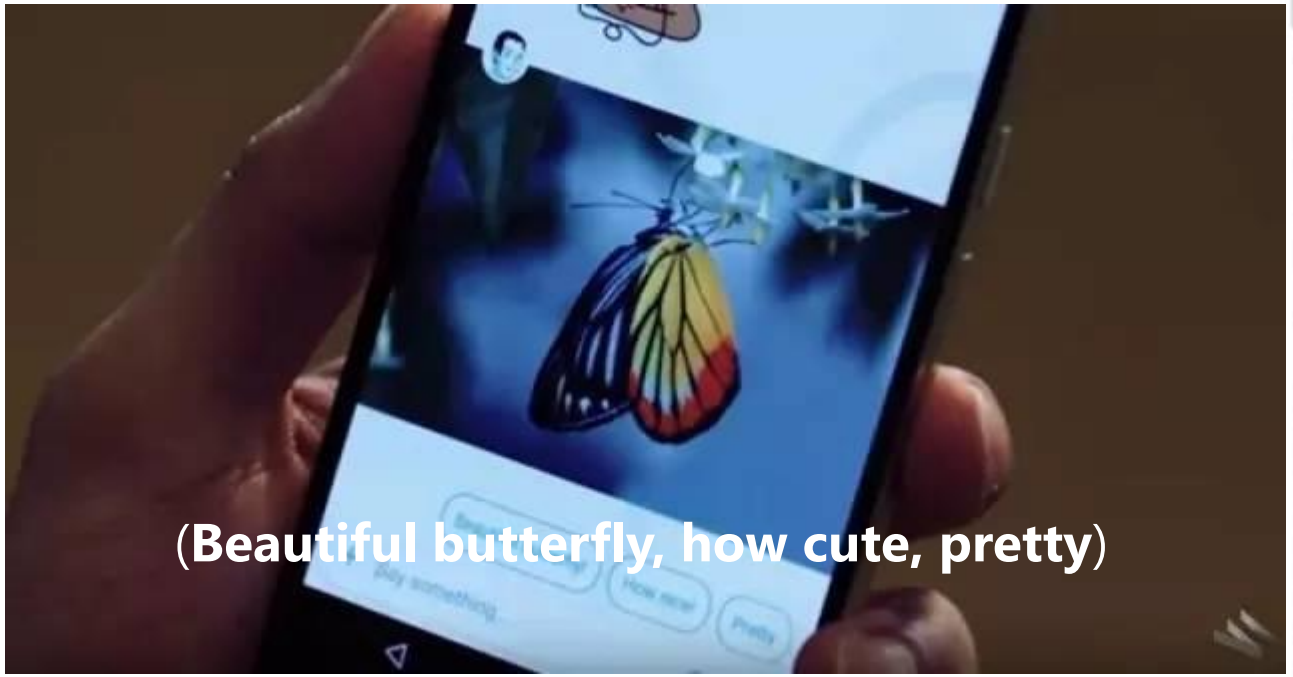
Assist the blind people

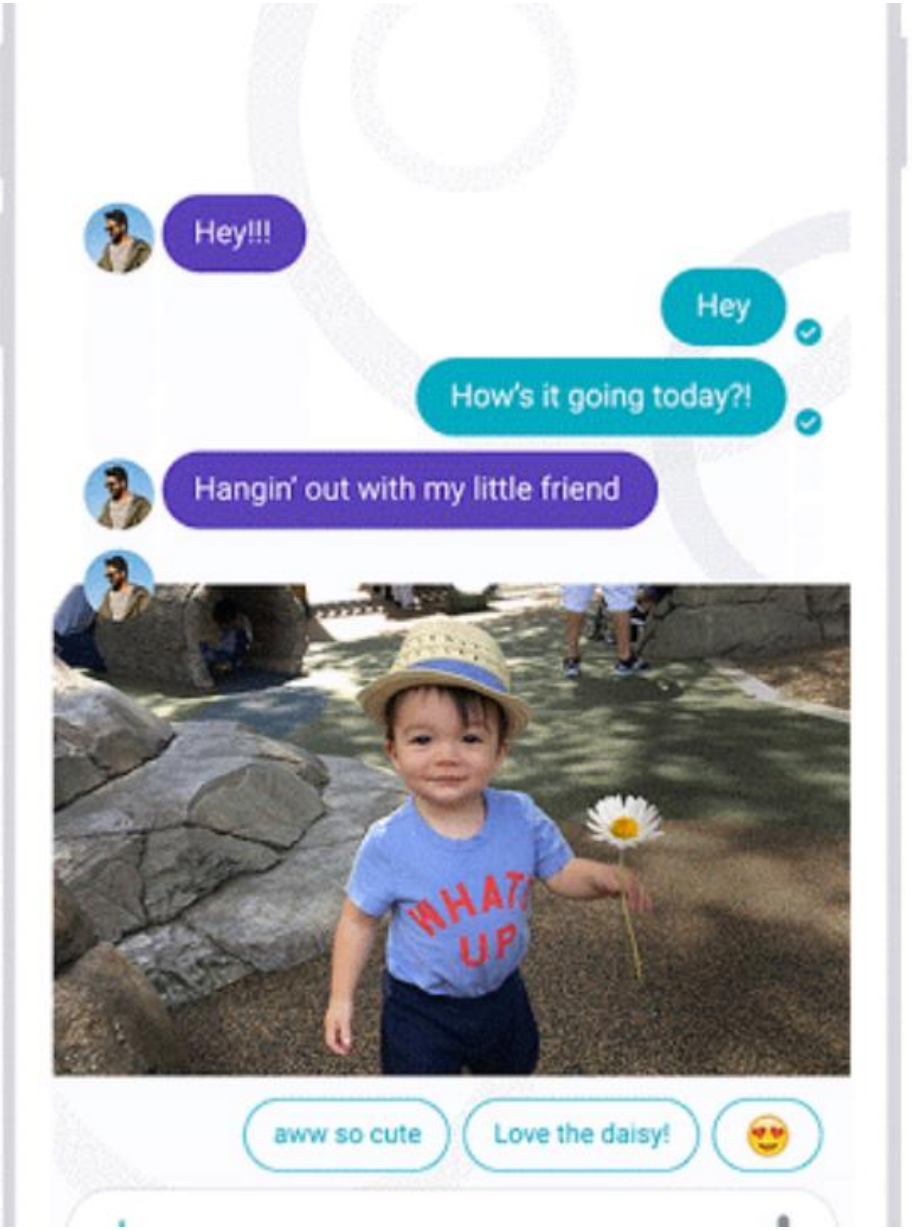
- Seeing AI: help the blind people to *see*

http://www.iqiyi.com/w_19rsyqc41h.html

http://blogs.microsoft.com/next/2016/03/30/decades-of-computer-vision-research-one-swiss-army-knife

Allo demo at Google I/O conference





Google Vision API – has tagging and landmark, no captioning yet

Outside Microsoft

Facebook announced an iOS app days after Build2016 – but more like keyword tagging rather than natural language description

Image may contain: pizza, food

Image may contain: tree, sky, outdoor

Image may contain: two people, smiling, sunglasses, sky, outdoor, water

From Captioning to Visual Question Answering

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

E.g., as a warning system for bicyclists, the system would interact with the cyclist, who might ask questions such as "*Are any other bikes going to pass me from the left?*" or "*Are there any runners close to me?*"

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

Image Question Answering

Answer:
dogs

Caption vs. QA: need reasoning

Reasoning in VQA:

Need to understand subtle relationships among multiple objects

Need to focus on the specific regions that are relevant to the answer.

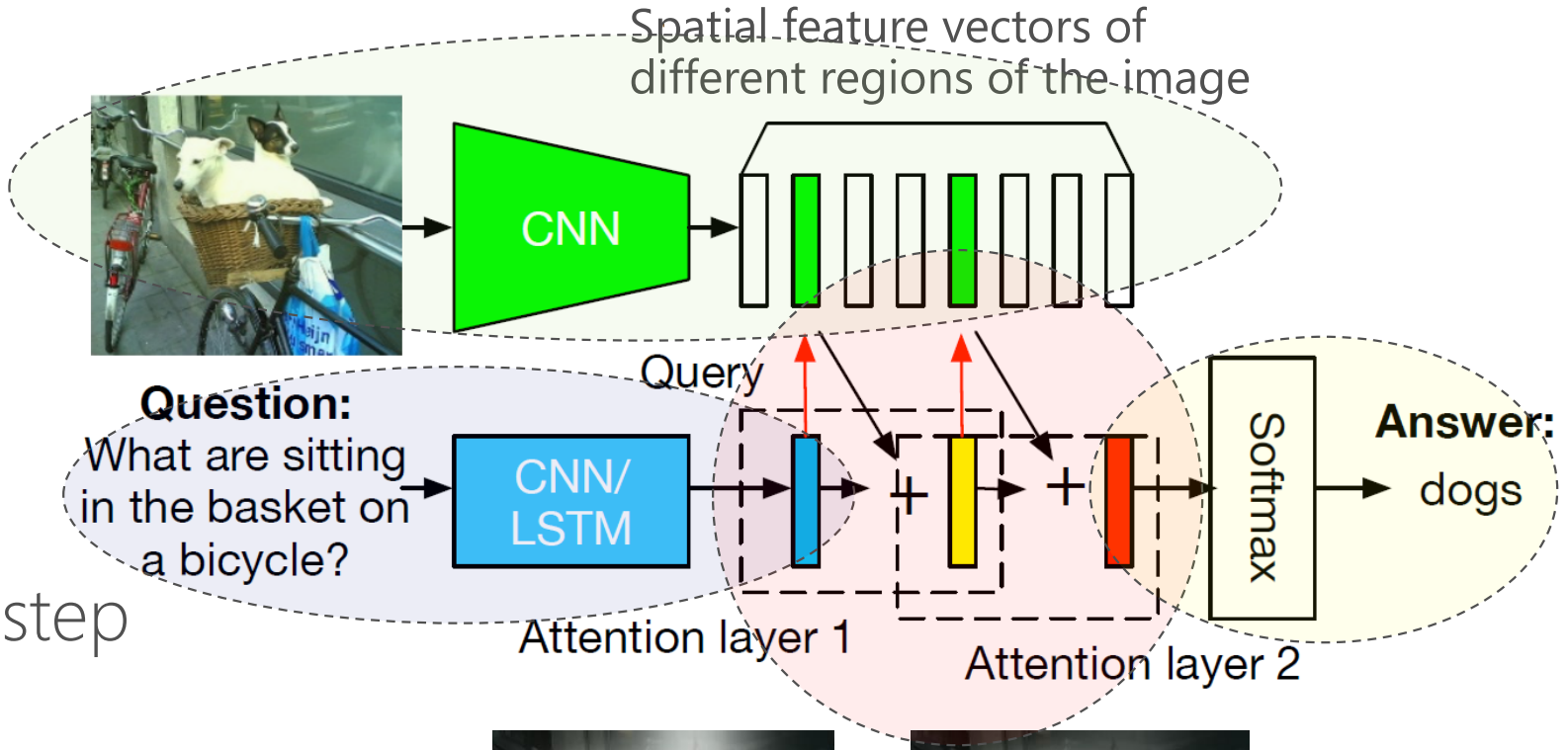
Question:
What are sitting in the basket on a bicycle?

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

Answer:
dogs

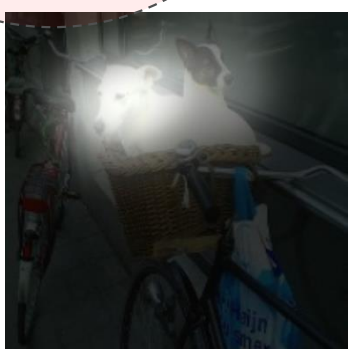
Stacked Attention Networks

[Yang, He, Gao, Deng, Smola, CVPR16]



SANs perform multi-step reasoning

1. Question model
2. Image model
3. Multi-level attention model
4. Answer predictor



1. The image model in the SAN

- Image Model

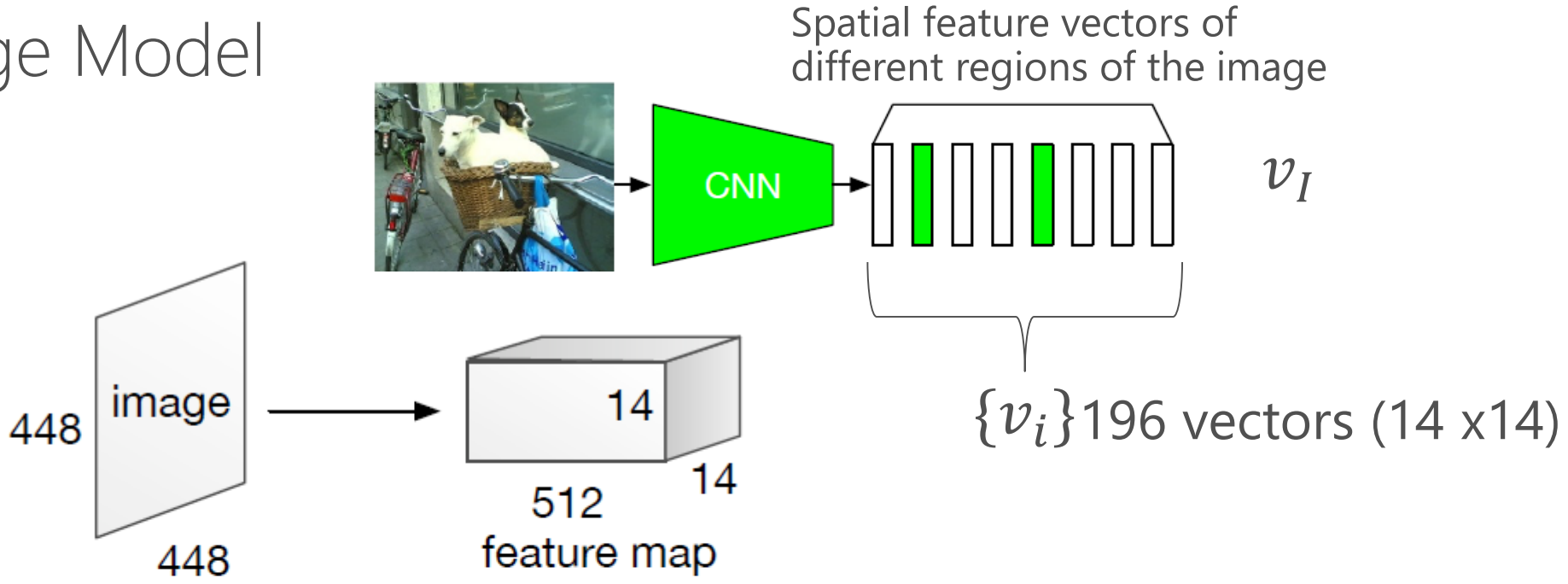


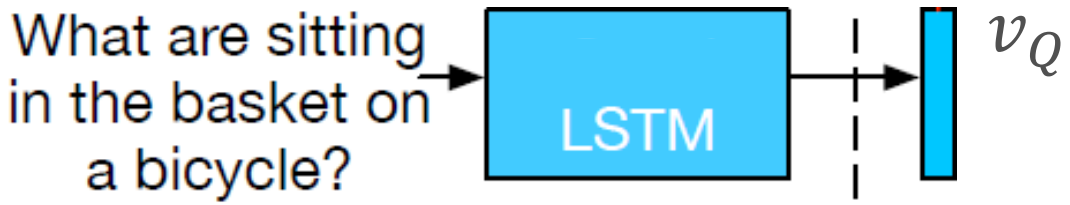
Figure 2: CNN based image model

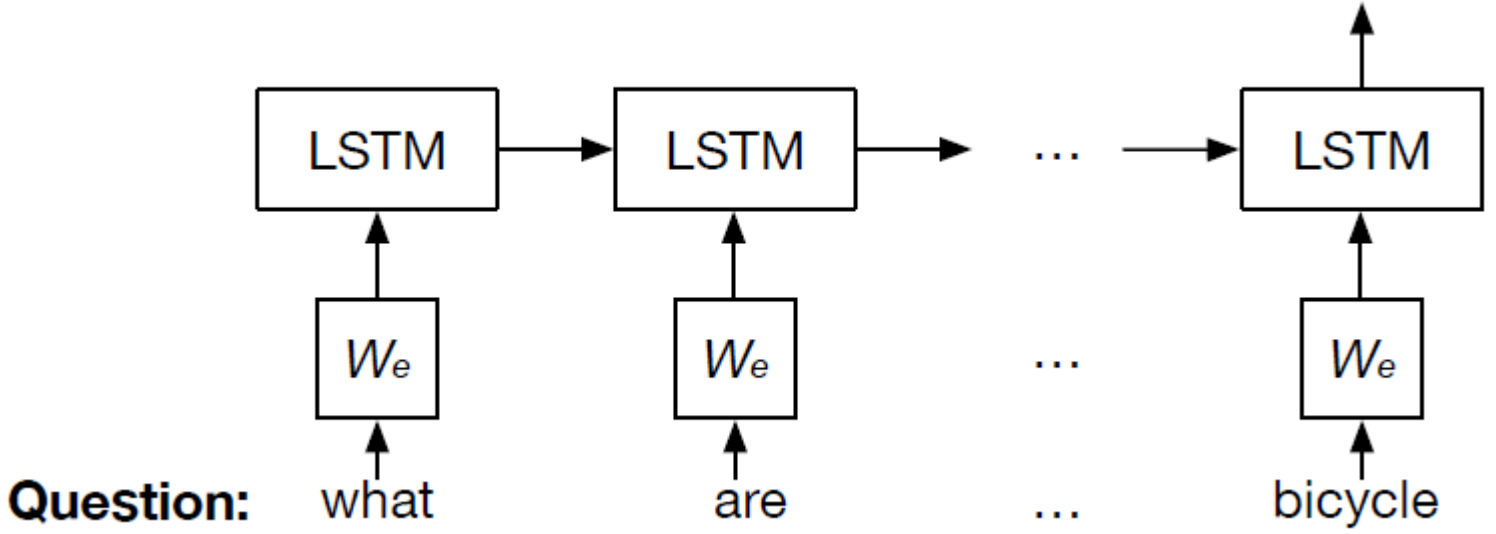
$$f_I = \text{CNN}_{vgg}(I). \qquad v_I = \tanh(W_I f_I + b_I)$$

2. The question model in the SAN

- Question Model

  Code the question into a vector using a LSTM

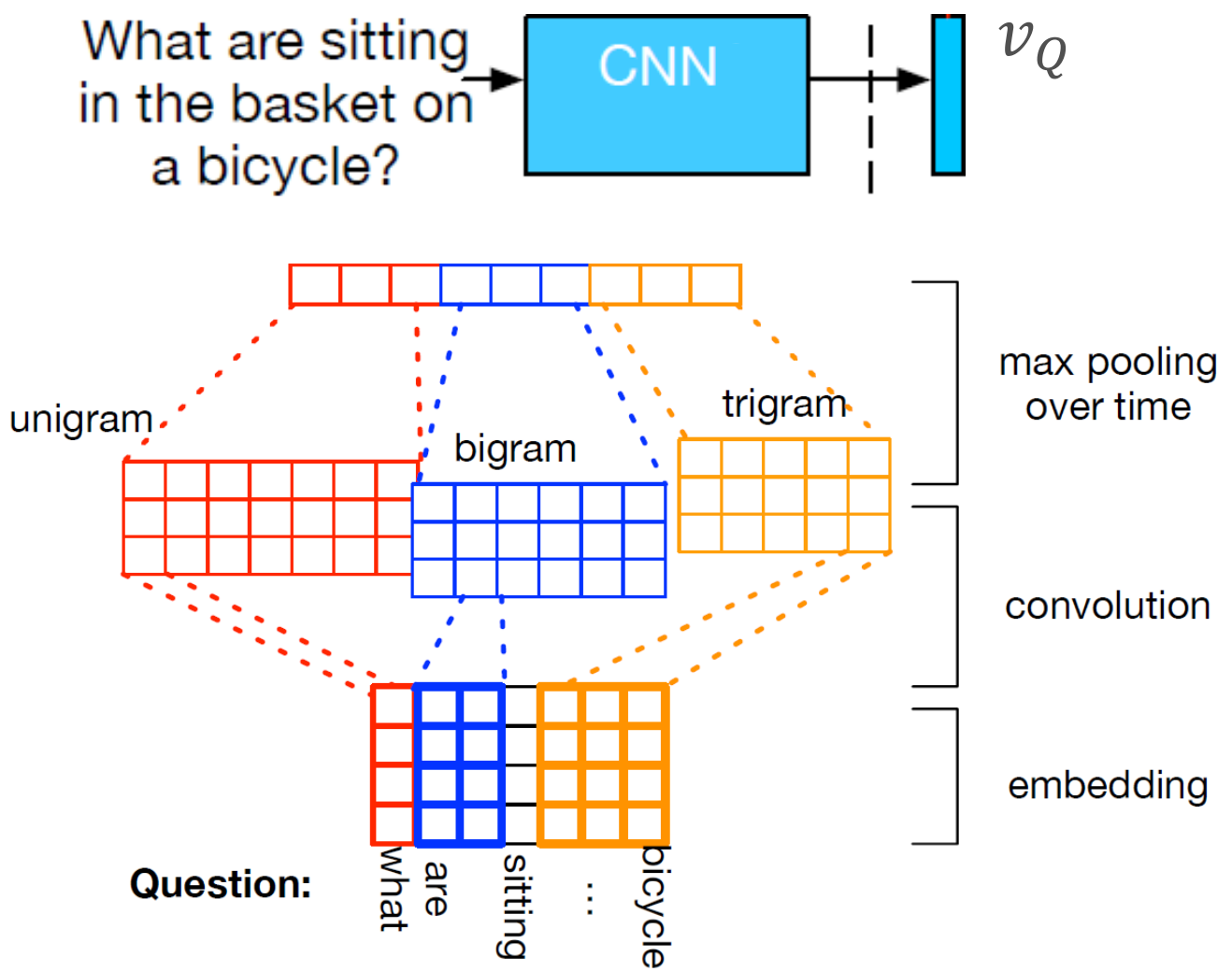




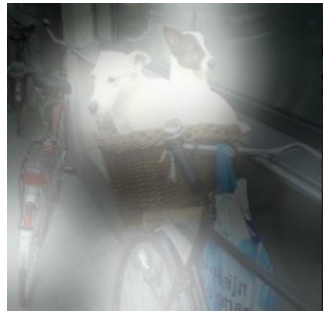
2. The question model in the SAN (alternative)

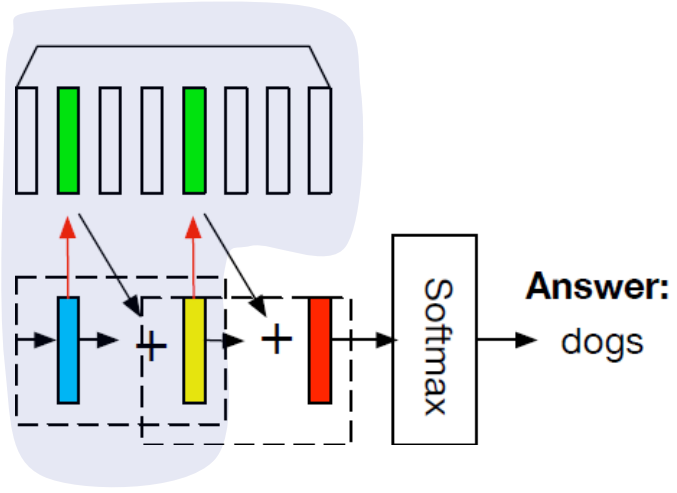
- Question Model

  Code the question into a vector using a CNN



3. SAN: Computing the 1st level attention





spatial image fea

$\{v_i\}$

Attention 1

v_Q

p_1 … p_{14}

attention map

$\{p_i\}$

p_{183} p_{196}

$$\tilde{v}_I = \sum_i p_i v_i$$

Multimodal Pooling (level 1)

$$u = \tilde{v}_I + v_Q$$

v_Q

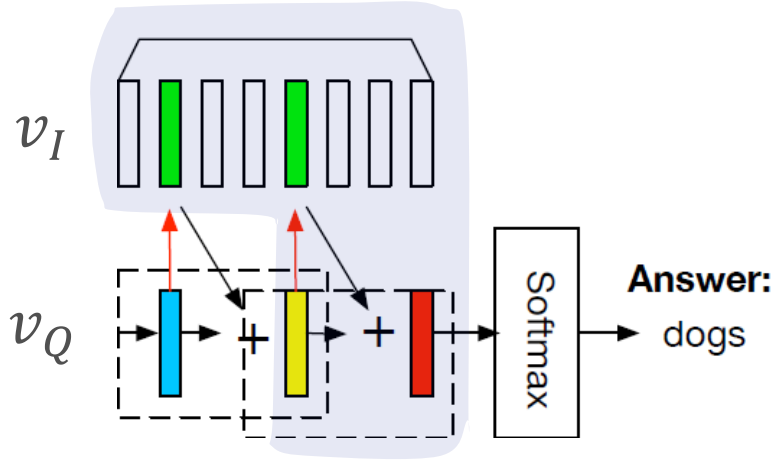
To the next attention level

3. SAN: Compute the 2nd level attention

spatial img fea

$\{v_i\}$

Attention 2



u

p_1 ... p_{14}

attention map

$\{p_i\}$

p_{183} p_{196}

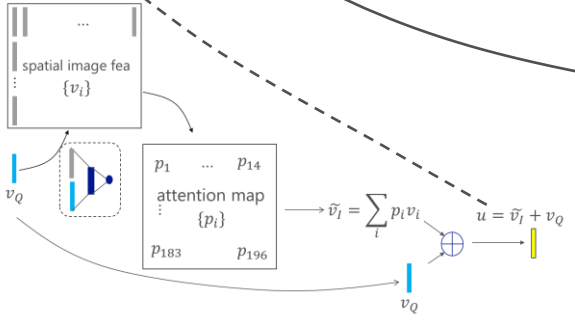
$$\tilde{v}_I^{(2)} = \sum_i p_i v_i$$

Multimodal Pooling (level 2)

$$u^{(2)} = \tilde{v}_I^{(2)} + u$$

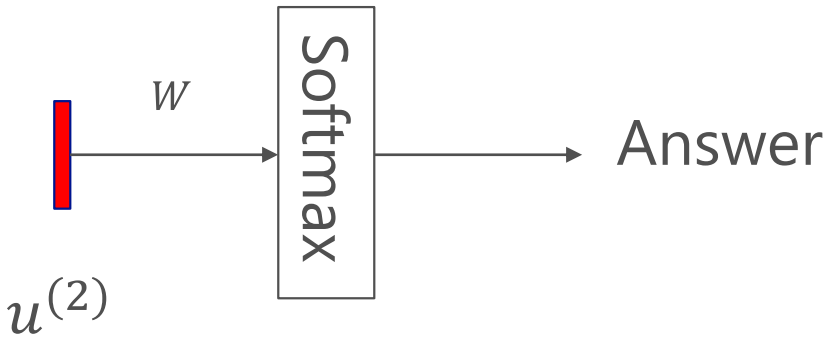
u

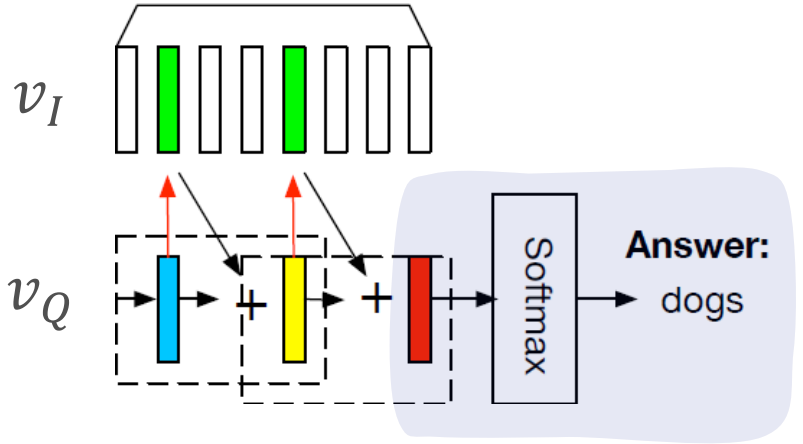
To the answer predictor



Query vector from the 1st level attention

4. Answer prediction





$$p_{ans} = softmax(W\, u^{(2)} + b)$$

$$ans^* = \underset{\{ans\}}{\text{argmax}}\{p_{ans}\}$$

Results

Methods	test-dev All	test-dev Yes/No	test-dev Number	test-dev Other	test-std 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**

Other:
Object
Color
Location
...

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

Big improvement on the VQA benchmark (and COCO-QA, DAQUAR)

Improvement is mainly in the *Other* category.

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

1st attention layer

2nd attention layer

Answer: **umbrella**

More examples:

(a) What are pulling a man on a wagon down on dirt road?
Answer: horses Prediction: horses

(b) What is the color of the box ?
Answer: red Prediction: red

(c) What next to the large umbrella attached to a table?
Answer: trees Prediction: tree

(d) How many people are going up the mountain with walking sticks?
Answer: four Prediction: four

(e) What is sitting on the handle bar of a bicycle?
Answer: bird Prediction: bird

(f) What is the color of the horns?
Answer: red Prediction: red

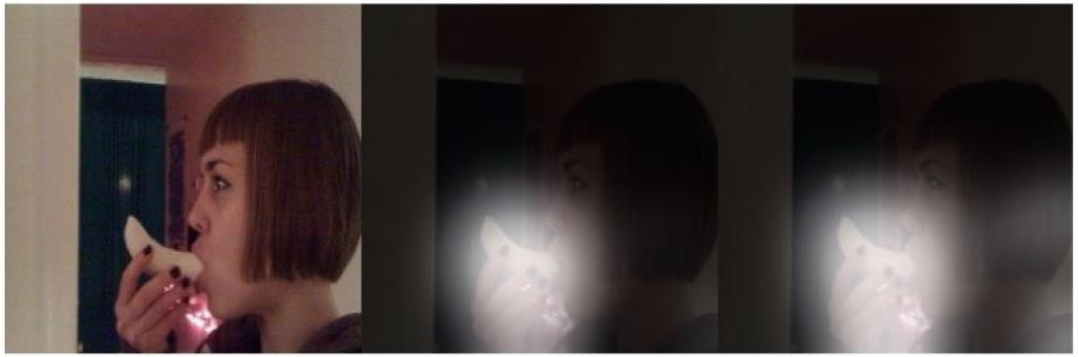
Original Image First Attention Layer Second Attention Layer Original Image First Attention Layer Second Attention Layer

Analysis (COCO-QA): 22% wrong attention, 42% wrong prediction, 31% ambiguous answer, 5% label error

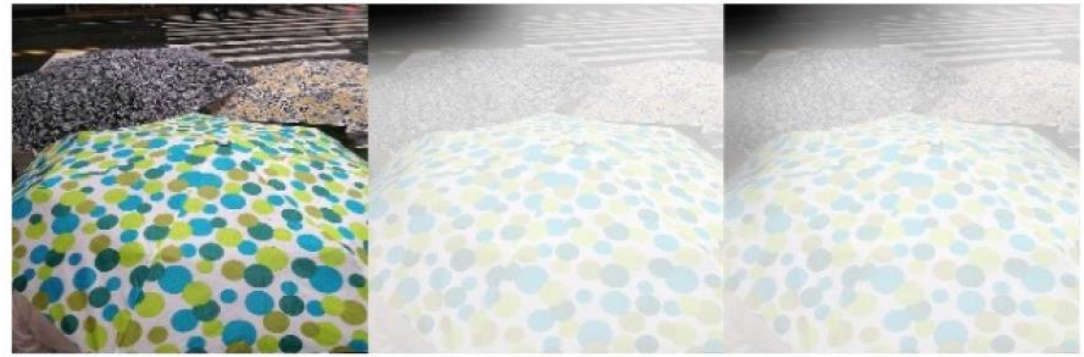
(a) What swim in the ocean near two large ferries?
Answer: ducks Prediction: boats

(b) What is the color of the shirt?
Answer: purple Prediction: green

(c) What is the young woman eating?
Answer: banana Prediction: donut



(d) How many umbrellas with various patterns?
Answer: three Prediction: two



(e) The very old looking what is on display?
Answer: pot Prediction: vase

(f) What are passing underneath the walkway bridge?
Answer: cars Prediction: trains

Original Image | First Attention Layer | Second Attention Layer | Original Image | First Attention Layer | Second Attention Layer

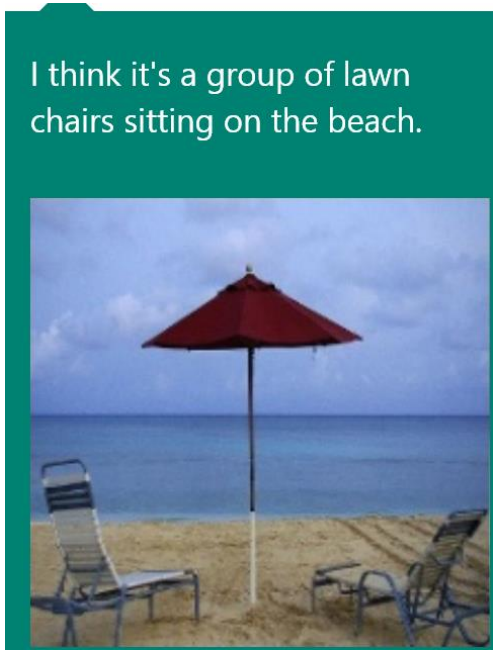
Deep Vision:

From captioning to QA, new challenges:

- subtle relationships among multiple objects
- focus on the specific region to infer the answer

From *generation* to *reasoning*

- use multi-level attention networks to infer the answer progressively
- perform visual grounded reasoning for VQA



http://captionBot.ai

Caption
and
QA

Q: what stands between the two lounge chairs on the beach?
A: **umbrella**

Go deeper?

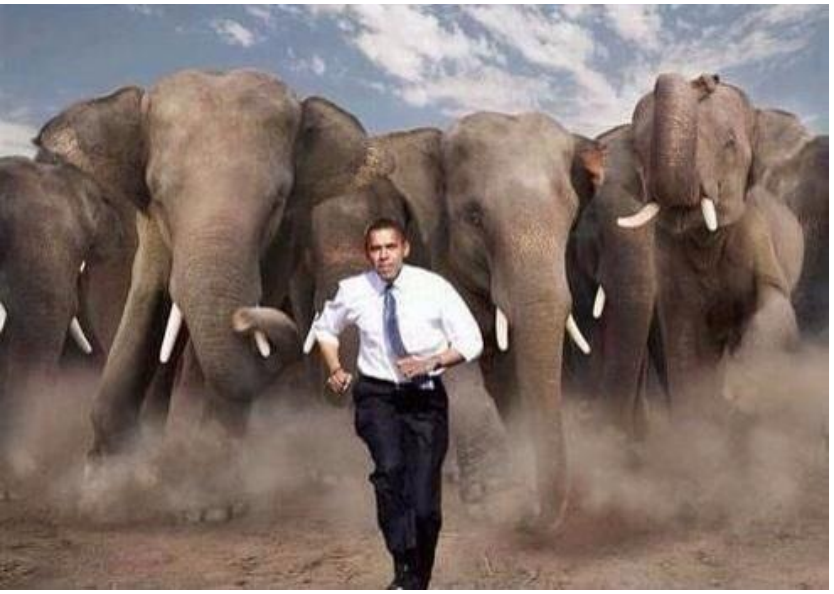


Image credit:
http://s122.photobucket.com/user/bmeupls/media/stampede.jpg.html

before:

➡ a herd of elephants standing next to a man

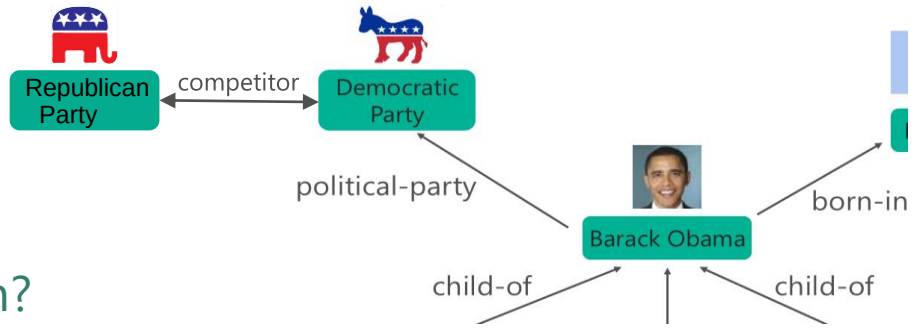
Now, + Entity:

➡ a herd of elephants standing next to Obama

Next, + knowledge & reasoning:

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

➡ Obama is chased by his republican competitors ☺



Knowledge Graph

Who is that person?
What are behind that man?
Why these elephants are chasing him?

Interim summary

Vision & language multimodal learning

- *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

- Image Captioning Service – CaptionBot http://CaptionBot.ai

Conclusions

- Exciting advances in NN and continuous representations
  - **Text processing & Knowledge reasoning**

- Looking forward
  - **Building an universal intelligence space**
    - Text, Knowledge, Reasoning, …
    - Sent2Vec (DSSM) http://aka.ms/sent2vec
  - **From component models to end-to-end solutions**

References

- Andreas, J., Rohrbach, M., Darrell, T., Klein, D., 2016. Neural Module Networks, CVPR
- Auli, M., Galley, M., Quirk, C. and Zweig, G., 2013. Joint language and translation modeling with recurrent neural networks. In EMNLP.
- Auli, M., and Gao, J., 2014. Decoder integration and expected bleu training for recurrent neural network language models. In ACL.
- Bahdanau, D., Cho, K., and Bengio, Y. 2015. Neural machine translation by joingly learning to align and translate, in ICLR 2015.
- Bejar, I., Chaffin, R. and Embretson, S. 1991. Cognitive and psychometric analysis of analogical problem solving. Recent research in psychology.
- Bengio, Y., 2009. Learning deep architectures for AI. Foundumental Trends Machine Learning, vol. 2.
- Bengio, Y., Courville, A., and Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE Trans. PAMI, vol. 38, pp. 1798-1828.
- Bengio, Y., Ducharme, R., and Vincent, P., 2000. A Neural Probabilistic Language Model, in NIPS.
- Berant, J., Chou, A., Frostig, R., Liang, P. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In EMNLP.
- Berant, J., and Liang, P. 2014. Semantic parsing via paraphrasing. In ACL.
- Bian, J., Gao, B., Liu, T. 2014. Knowledge-Powered Deep Learning for Word Embedding. In ECML.
- Blei, D., Ng, A., and Jordan M. 2001. Latent dirichlet allocation. In NIPS.
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J. and Yakhnenko, O. 2013. Translating Embeddings for Modeling Multi-relational Data. In NIPS.
- Bordes, A., Chopra, S., and Weston, J. 2014. Question answering with subgraph embeddings. In EMNLP.
- Bordes, A., Glorot, X., Weston, J. and Bengio Y. 2012. Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing. In AISTATS.
- Brown, P., deSouza, P. Mercer, R., Della Pietra, V., and Lai, J. 1992. Class-based n-gram models of natural language. Computational Linguistics 18 (4).
- Chandar, A. P. S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V., and Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In NIPS.
- Chang, K., Yih, W., and Meek, C. 2013. Multi-Relational Latent Semantic Analysis. In EMNLP.
- Chang, K., Yih, W., Yang, B., and Meek, C. 2014. Typed Tensor Decomposition of Knowledge Bases for Relation Extraction. In EMNLP.
- Collobert, R., and Weston, J. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In ICML.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P., 2011. Natural language processing (almost) from scratch. in JMLR, vol. 12.
- Cui, L., Zhang, D., Liu, S., Chen, Q., Li, M., Zhou, M., and Yang, M. (2014). Learning topic representation for SMT with neural networks. In ACL.
- Dahl, G., Yu, D., Deng, L., and Acero, 2012. A. Context-dependent, pre-trained deep neural networks for large vocabulary speech recognition, IEEE Trans. Audio, Speech, & Language Proc., Vol. 20 (1), pp. 30-42.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T., and Harshman, R. 1990. Indexing by latent semantic analysis. J. American Society for Information Science, 41(6): 391-407
- Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., and Mitchell, M., 2015. Language Models for Image Captioning: The Quirks and What Works, ACL

References

- Deng, L., He, X., Gao, J., 2013. Deep stacking networks for information retrieval, ICASSP
- Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., and Hinton, G., 2010. Binary Coding of Speech Spectrograms Using a Deep Auto-encoder, in Interspeech.
- Deng, L., Tur, G, He, X, and Hakkani-Tur, D. 2012. Use of kernel deep convex networks and end-to-end learning for spoken language understanding, Proc. IEEE Workshop on Spoken Language Technologies.
- Deng, L., Yu, D. and Acero, A. 2006. Structured speech modeling, IEEE Trans. on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1492-1504.
- Deng, L., Yu, D., and Platt, J. 2012. Scalable stacking and learning for building deep architectures, Proc. ICASSP.
- Deng, L. and Yu, D. 2014. Deeping learning methods and applications. Foundations and Trends in Signal Processing 7:3-4.
- Deoras, A., and Sarikaya, R., 2013. Deep belief network based semantic taggers for spoken language understanding, in INTERSPEECH.
- Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J., 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation, ACL.
- Duh, K. 2014. Deep learning for natural language processing and machine translation. Tutorial. 2014.
- Duh, K., Neubig, G., Sudoh, K., and Tsukada, H. (2013). Adaptation data selection using neural language models: Experiments in machine translation. In ACL.
- Fader, A., Zettlemoyer, L., and Etzioni, O. 2013. Paraphrase-driven learning for open question answering. In ACL.
- Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., Zitnick, L., Zweig, G., "From Captions to Visual Concepts and Back," arXiv:1411.4952
- Faruqui, M. and Dyer, C. (2014). Improving vector space word representations using multilingual correlation. In EACL.
- Faruqui, M., Dodge, J., Jauhar, S., Dyer, C., Hovy, E., Smith, N. 2015. Retrofitting Word Vectors to Semantic Lexicons. In NAACL-HLT.
- Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., Smith, N. 2015. Sparse Overcomplete Word Vector Representations. In ACL.
- Firth, J. R. 1957. *Papers in Linguistics 1934–1951*, Oxford University Press, 1957
- Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T., 2013. DeViSE: A Deep Visual-Semantic Embedding Model, Proc. NIPS.
- Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F. 2013. Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases. In WWW.
- Gao, J., He, X., Yih, W-t., and Deng, L. 2014a. Learning continuous phrase representations for translation modeling. In ACL.
- Gao, J., He, X., and Nie, J-Y. 2010. Clickthrough-based translation models for web search: from word models to phrase models. In CIKM.
- Gao, J., Pantel, P., Gamon, M., He, X., Deng, L., and Shen, Y. 2014b. Modeling interestingness with deep neural networks. In EMNLP
- Gao, J., Toutanova, K., Yih., W-T. 2011. Clickthrough-based latent semantic models for web search. In SIGIR.

References

- Gao, J., Yuan, W., Li, X., Deng, K., and Nie, J-Y. 2009. Smoothing clickthrough data for web search ranking. In SIGIR.
- Gao, J., and He, X. 2013. Training MRF-based translation models using gradient ascent. In NAACL-HLT.
- Getoor, L., and Taskar, B. editors. 2007. Introduction to Statistical Relational Learning. The MIT Press.
- Graves, A., Jaitly, N., and Mohamed, A., 2013a. Hybrid speech recognition with deep bidirectional LSTM, Proc. ASRU.
- Graves, A., Mohamed, A., and Hinton, G., 2013. Speech recognition with deep recurrent neural networks, Proc. ICASSP.
- He, J., Chen, J., He, X., Gao, J., Li, L., Deng, L., Ostendorf, M., 2015 Deep Reinforcement Learning with an Action Space Defined by Natural Language, arXiv:1511.04636 (to appear on EMNLP16)
- He, X. and Deng, L., 2013. Speech-Centric Information Processing: An Optimization-Oriented Approach, in Proceedings of the IEEE.
- He, X. and Deng, L., 2012. Maximum Expected BLEU Training of Phrase and Lexicon Translation Models , ACL.
- He, X., Deng, L., and Chou, W., 2008. Discriminative learning in sequential pattern recognition, Sept. IEEE Sig. Proc. Mag.
- Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. In ACL.
- Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B., 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97.
- Hinton, G., Osindero, S., and The, Y-W. 2006. A fast learning algorithm for deep belief nets. *Neural Computation*, 18: 1527-1554.
- Hinton, G., and Salakhutdinov, R., 2010. Discovering binary codes for documents by learning deep generative models. Topics in Cognitive Science.
- Hu, Y., Auli, M., Gao, Q., and Gao, J. 2014. Minimum translation modeling with recurrent neural networks. In EACL.
- Huang, E., Socher, R., Manning, C, and Ng, A. 2012. Improving word representations via global context and multiple word prototypes, Proc. ACL.
- Huang, P., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In CIKM.
- Hutchinson, B., Deng, L., and Yu, D., 2012. A deep architecture with bilinear modeling of hidden representations: Applications to phonetic recognition, Proc. ICASSP.
- Hutchinson, B., Deng, L., and Yu, D., 2013. Tensor deep stacking networks, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, pp. 1944 - 1957.
- Jansen, P., Surdeanu, M., Clark, P. 2014. Discourse Complements Lexical Semantics for Non-factoid Answer Reranking. In ACL.
- Jurgens, D., Mohammad, S., Turney, P. and Holyoak, K. 2012. SemEval-2012 Task 2: Measuring degrees of relational similarity. In SemEval.
- Kafle, K., Kanan, C., 2016. Answer-Type Prediction for Visual Question Answering, CVPR
- Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models., in EMNLLP
- Kiros, R., Zemel, R., and Salakhutdinov, R. 2013. Multimodal Neural Language Models, Proc. NIPS Deep Learning Workshop.
- Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In COLING.
- Kocisky, T., Hermann, K. M., and Blunsom, P. (2014). Learning bilingual word representations by marginalizing alignments. In ACL.
- Koehn, P. 2009. Statistical Machine Translation. Cambridge University Press.

References

- Krizhevsky, A., Sutskever, I, and Hinton, G., 2012. ImageNet Classification with Deep Convolutional Neural Networks, NIPS.
- Landauer. T., 2002. On the computational basis of learning and cognition: Arguments from LSA. Psychology of Learning and Motivation, 41:43–84.
- Lao, N., Mitchell, T., and Cohen, W. 2011. Random walk inference and learning in a large scale knowledge base. In EMNLP.
- Lauly, S., Boulanger, A., and Larochelle, H. (2013). Learning multilingual word representations using a bag-of-words autoencoder. In NIPS.
- Le, H-S, Oparin, I., Allauzen, A., Gauvain, J-L., Yvon, F., 2013. Structured output layer neural network language models for speech recognition, IEEE Transactions on Audio, Speech and Language Processing.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. Gradient-based learning applied to document recognition, Proceedings of the IEEE, Vol. 86, pp. 2278-2324.
- Levy, O., and Goldberg, Y. 2014. Linguistic Regularities in Sparse and Explicit Word Representations. In CoNLL.
- Levy, O., and Goldberg, Y. 2014. Neural Word Embeddings as Implicit Matrix Factorization. In NIPS.
- Li, P., Hastie, T., and Church, K.. 2006. Very sparse random projections, in Proc. SIGKDD.
- Li, P., Liu, Y., and Sun, M. (2013). Recursive autoencoders for ITG-based translation. In EMNLP.
- Li, P., Liu, Y., Sun, M., Izuha, T., and Zhang, D. (2014b). A neural reordering model for phrase-based translation. In COLING.
- Liu, S., Yang, N., Li, M., and Zhou, M. (2014). A recursive recurrent neural network for statistical machine translation. In ACL.
- Liu, X., Gao, J., He, X., Deng, L., Duh, K., Wang, Y., 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval, NAACL
- Liu, L., Watanabe, T., Sumita, E., and Zhao, T. (2013). Additive neural networks for statistical machine translation. In ACL.
- Lu, S., Chen, Z., and Xu, B. (2014). Learning new semi-supervised deep auto-encoder features for statistical machine translation. In ACL.
- Maskey, S., and Zhou, B. 2012. Unsupervised deep belief feature for speech translation, in ICASSP.
- Mesnil, G., He, X., Deng, L., and Bengio, Y., 2013. Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding, in Interspeech.
- Mikolov, T., Kombrink, S., Burget, L., Cernocky, J., Khudanpur, S. 2011. Extensions of recurrent neural network based language model. In ICASSP.
- Mikolov, T. 2012. Statistical Language Models based on Neural Networks, Ph.D. thesis, Brno University of Technology.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficient estimation of word representations in vector space, Proc. ICLR.
- Mikolov, T., Kombrink,. S., Burget, L., Cernocky, J., Khudanpur, S., 2011. Extensions of Recurrent Neural Network LM. ICASSP.
- Mikolov, T., Yih, W., Zweig, G., 2013. Linguistic Regularities in Continuous Space Word Representations. In NAACL-HLT.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In NIPS.
- Mnih, A., Kavukcuoglu, K. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In NIPS.
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing Atari with Deep Reinforcement Learning, NIPS

References

- Mohamed, A., Yu, D., and Deng, L. 2010. Investigation of full-sequence training of deep belief networks for speech recognition, Proc. Interspeech.
- Mohammad, S., Dorr, Bonnie., and Hirst, G. 2008. Computing word pair antonymy. In EMNLP.
- Narasimhan, K., Kulkarni, T., Barzilay, R., 2015. Language Understanding for Text-based Games Using Deep Reinforcement Learning. EMNLP
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. 2011. Multimodal deep learning, Proc. ICML.
- Nickel, M., Tresp, V., and Kriegel, H. 2011. A three-way model for collective learning on multi-relational data. In ICML.
- Niehues, J., Waibel, A. 2013. Continuous space language models using Restricted Boltzmann Machines. In IWLT.
- Noh, H., Seo, P., Han, B., 2016. Image Question Answering Using Convolutional Neural Network With Dynamic Parameter Prediction, CVPR
- Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., Ward R., 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval, IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (4), 694-707
- Pennington, J., Socher, R., Manning, C. 2014. Glove: Global Vectors for Word Representation. In EMNLP.
- Reddy, S., Lapata, M., and Steedman, M. 2014. Large-scale semantic parsing without question-answer pairs. Transactions of the Association for Computational Linguistics (TACL).
- Sainath, T., Mohamed, A., Kingsbury, B., and Ramabhadran, B. 2013. Convolutional neural networks for LVCSR, Proc. ICASSP.
- Salakhutdinov R., and Hinton, G., 2007 Semantic hashing. in Proc. SIGIR Workshop Information Retrieval and Applications of Graphical Models
- Salton, G. and McGill, M. 1983. Introduction to Modern Information Retrieval. McGraw Hill.
- Sarikaya, R., Hinton, G., and Ramabhadran, B., 2011. Deep belief nets for natural language call-routing, in Proceedings of the ICASSP.
- Schwenk, H. 2012. Continuous space translation models for phrase-based statistical machine translation, in COLING.
- Schwenk, H., Rousseau, A., and Attik, M., 2012. Large, pruned or continuous space language models on a gpu for statistical machine translation, in NAACL-HLT 2012 Workshop.
- Seide, F., Li, G., and Yu, D. 2011. Conversational speech transcription using context-dependent deep neural networks, Proc. Interspeech
- Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. 2014. Learning Semantic Representations Using Convolutional Neural Networks for Web Search, in Proceedings of WWW.
- Shen, Y., He, X., Gao, J., Deng, L., and Mesnil, G. 2014. A convolutional latent semantic model for web search. CIKM
- Shih, K., Singh, S., Hoiem, D., 2016. Where to Look: Focus Regions for Visual Question Answering, CVPR
- Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of Go with deep neural networks and tree search, Nature
- Simonyan, K., Zisserman, A., 2015 Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015
- Socher, R., Chen, D., Manning, C., and Ng, A. 2013. Reasoning With Neural Tensor Networks For Knowledge Base Completion. In NIPS.
- Socher, R., Huval, B., Manning, C., Ng, A., 2012. Semantic compositionality through recursive matrix-vector spaces. In EMNLP.
- Socher, R., Lin, C., Ng, A., and Manning, C. 2011. Learning continuous phrase representations and syntactic parsing with recursive neural networks, Proc. ICML.

References

- Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng A., and Potts. C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, Proc. EMNLP
- Son, L. H., Allauzen, A., and Yvon, F. (2012). Continuous space translation models with neural networks. In NAACL.
- Song, X. He, X., Gao. J., and Deng, L. 2014. Unsupervised Learning of Word Semantic Embedding using the Deep Structured Semantic Model. MSR Tech Report.
- Song, Y., Wang, H., and He, X., 2014. Adapting Deep RankNet for Personalized Search. Proc. WSDM.
- Songyot, T. and Chiang, D. (2014). Improving word alignment using word similarity. In EMNLP.
- Sundermeyer, M., Alkhouli, T., Wuebker, J., and Ney, H. (2014). Translation modeling with bidirectional recurrent neural networks, in EMNLP.
- Sutton, R., Barto, A., 1998. Reinforcement Learning: An Introduction. MIT Press.
- Tamura, A., Watanabe, T., and Sumita, E. (2014). Recurrent neural networks for word alignment model. In ACL.
- Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S., 2016. MovieQA: Understanding Stories in Movies Through Question-Answering, CVPR
- Tran, K. M., Bisazza, A., and Monz, C. (2014). Word translation prediction for morphologically rich languages with bilingual neural networks. In EMNLP.
- Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., Sienkiewicz, C., "Rich Image Captioning in the Wild," DeepVision, CVPR 2016
- Tur, G., Deng, L., Hakkani-Tur, D., and He, X., 2012. Towards Deeper Understanding Deep Convex Networks for Semantic Utterance Classification, in ICASSP.
- Turney P. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In COLING. Songyot, T. and Chiang, D. (2014). Improving word alignment using word similarity. In EMNLP.
- Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. 2013. Decoding with large-scale neural language models improves translation. In EMNLP.
- Wang, H., He, X., Chang, M., Song, Y., White, R., Chu, W., 2013. Personalized ranking model adaptation for web search, SIGIR Wang, Z., Zhang, J., Feng, J., Chen, Z. 2014. Knowledge Graph and Text Jointly Embedding. In EMNLP.
- Watkins, C., and Dayan, P., 1992. Q-learning. Machine Learning
- Wright, S., Kanevsky, D., Deng, L., He, X., Heigold, G., and Li, H., 2013. Optimization Algorithms and Applications for Speech and Language Processing, in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 11.
- Wu, Q., Wang, P., Shen, C., Dick, A., Hengel, A., 2016. Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge From External Sources, CVPR
- Wu, H., Dong, D., Hu, X., Yu, D., He, W., Wu, H., Wang, H., and Liu, T. (2014a). Improve statistical machine translation with context-sensitive bilingual semantic embedding model. In EMNLP.
- Wu, Y., Watanabe, T., and Hori, C. (2014b). Recurrent neural network-based tuple sequence model for machine translation. In COLING.
- Xu, C., Bai, Y., Bian, J., Gao, B., Wang, G., Liu, X., Liu, T. 2014. RC-NET: A General Framework for Incorporating Knowledge into Word Representations. In CIKM.
- Yang, B., Yih, W., He, X., Gao, J., and Deng L. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In ICLR.
- Yang, N., Liu, S., Li, M., Zhou, M., and Yu, N. 2013. Word alignment modeling with context dependent deep neural network. In ACL.
- Yang, Y., Chang, M. 2015. S-MART: Novel Tree-based Structured Learning Algorithms Applied to Tweet Entity Linking. In ACL.
- Yao, K., Zweig, G., Hwang, M-Y. , Shi, Y., Yu, D., 2013. Recurrent neural networks for language understanding, submitted to Interspeech.

References

- Yao, X., Van Durme, B. 2014. Information Extraction over Structured Data: Question Answering with Freebase. In ACL.
- Yann, D., Tur, G., Hakkani-Tur, D., Heck, L., 2014. Zero-Shot Learning and Clustering for Semantic Utterance Classification Using Deep Learning. In ICLR
- Yogatama, D., Faruqui, M., Dyer, C., Smith, N. 2015. LearningWord Representations with Hierarchical Sparse Coding. In ICML.
- Yih, W., Toutanova, K., Platt, J., and Meek, C. 2011. Learning discriminative projections for text similarity measures. In CoNLL.
- Yih, W., Zweig, G., Platt, J. 2012. Polarity Inducing Latent Semantic Analysis. In EMNLP-CoNLL.
- Yih, W., Chang, M., Meek, C., Pastusiak, A. 2013. Question Answering Using Enhanced Lexical Semantic Models. In ACL.
- Yih, W., He, X., Meek, C. 2014. Semantic Parsing for Single-Relation Question Answering. In ACL.
- Yih, W., Chang, M., He, X., Gao, J. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base, In ACL.
- Zeiler, M. and Fergus, R. 2013. Visualizing and understanding convolutional networks, arXiv:1311.2901, pp. 1-11.
- Zhang, J., Liu, S., Li, M., Zhou, M., and Zong, C. (2014). Bilingually-constrained phrase embeddings for machine translation. In ACL.
- Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L., 2016. Visual7W: Grounded Question Answering in Images, CVPR
- Zou, W. Y., Socher, R., Cer, D., and Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In EMNLP.