Module 6

Vision-Language Multimodal Intelligence

Vision-Language Multimodal Intelligence

- Image Captioning
- Visual Question answering

Humans learn to process text, image, and knowledge jointly

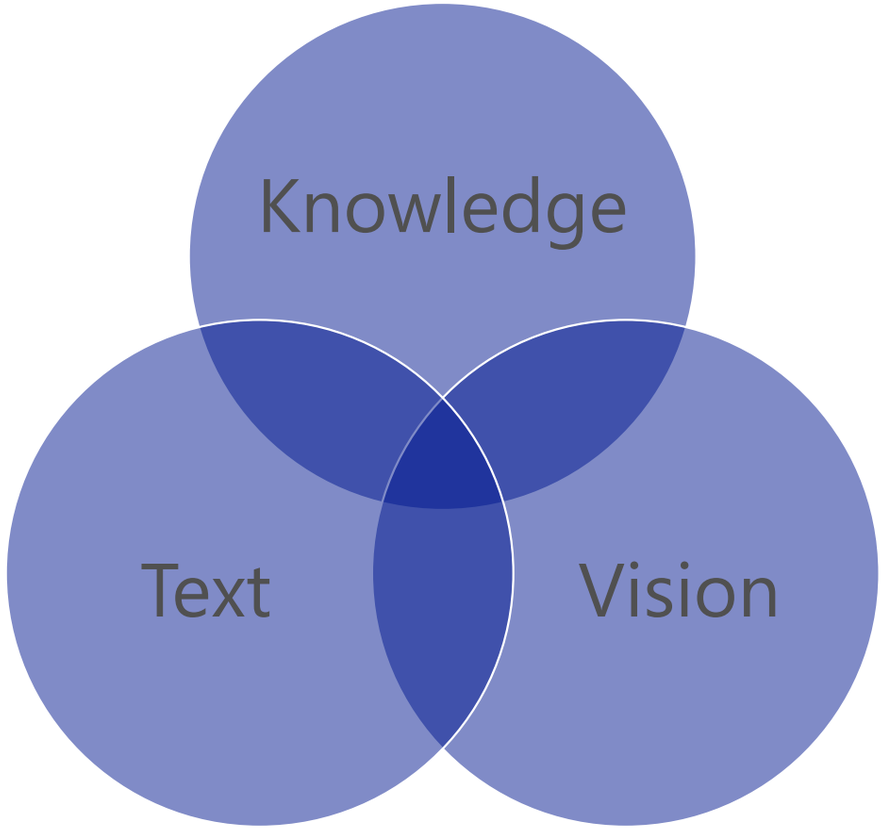


Image Captioning (one step from perception to cognition)

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

a man holding a tennis racquet on a tennis court

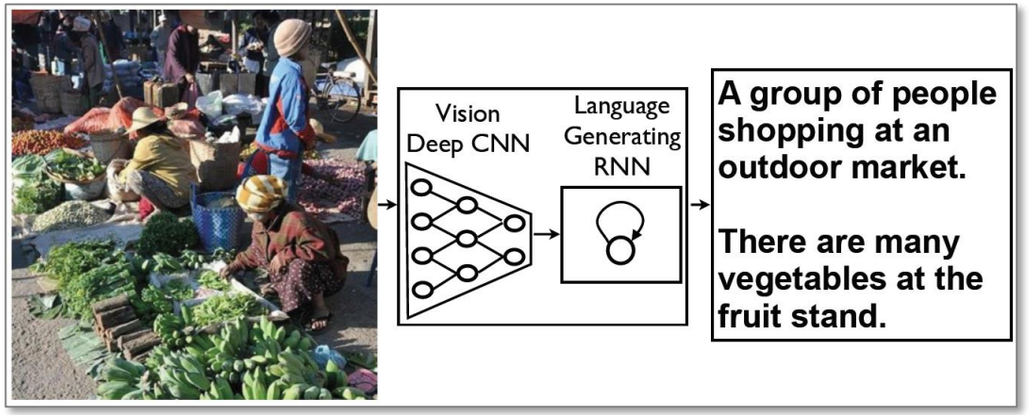
the man is on the tennis court playing a game

-- Let's do a Turing Test!

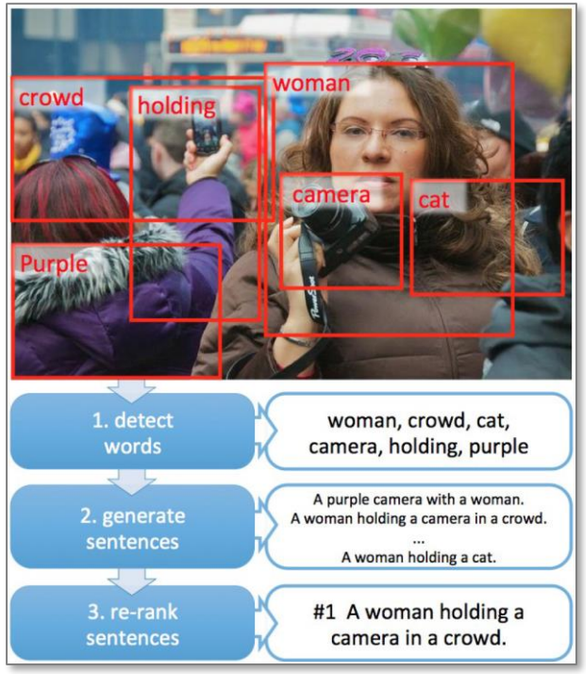
Two major paradigms

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 Image captioning

- Image word detection
  - Deep-learned model to detect key concepts in the image
- Language model generates caption candidates
  - Maxent language model (MELM) conditional on words detected from the image
- Deep multi-modal semantic model re-ranking
  - Hypothetical captions re-ranked by deep-learned multimodal similarity model (DMSM) looking at the entire image

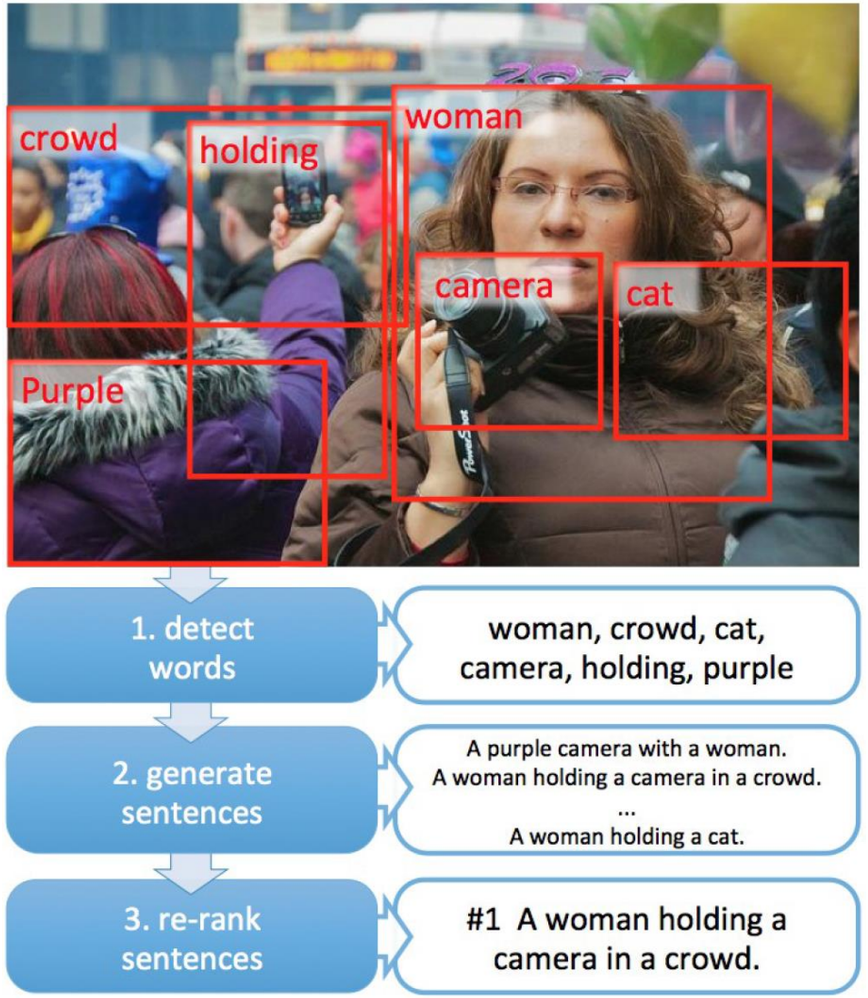


Figure 1. An illustrative example of our pipeline.

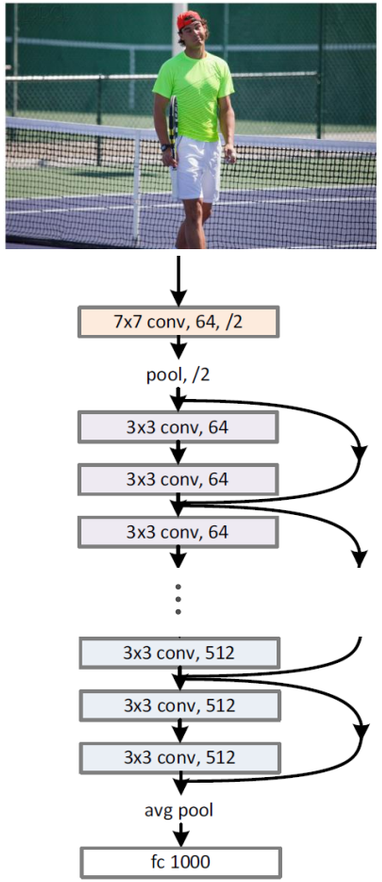
[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "From Captions to Visual Concepts and Back," CVPR, June 2015]

Deep CNN for visual concepts detection

ResNet

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

Trained on multiple GPUs



[He, Zhang, Ren, Sun, CVPR2015]

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

MaxEnt LM (MELM) for modeling language

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

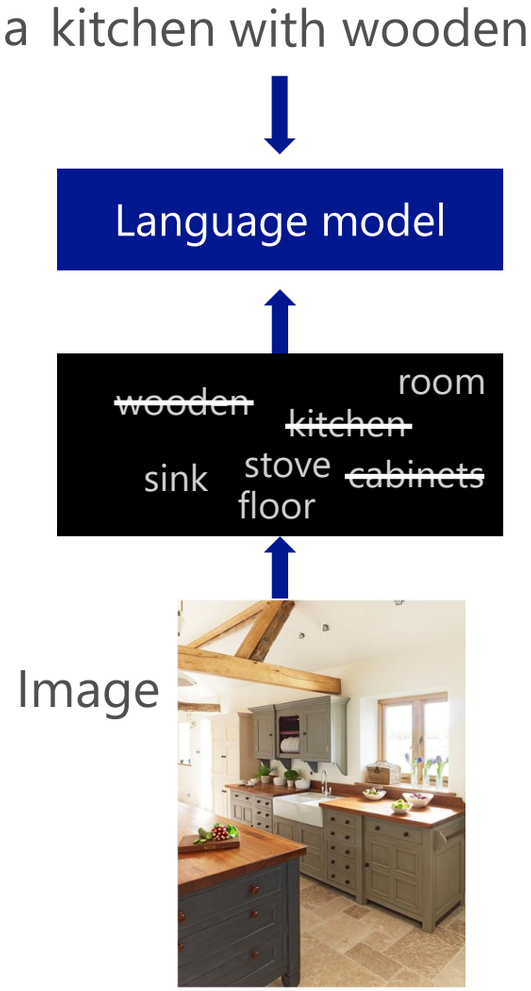
Feature	Type	Definition	Description
Attribute	0/1	$\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	Predicted word is in the attribute set, i.e. has been visually detected and not yet used.
N-gram+	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	N-gram ending in 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



cabinets

MaxEnt LM

$p(cabinets|with\ wooden)$

a kitchen with wooden cabinets

Beam search to generate 500 candidates

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

...

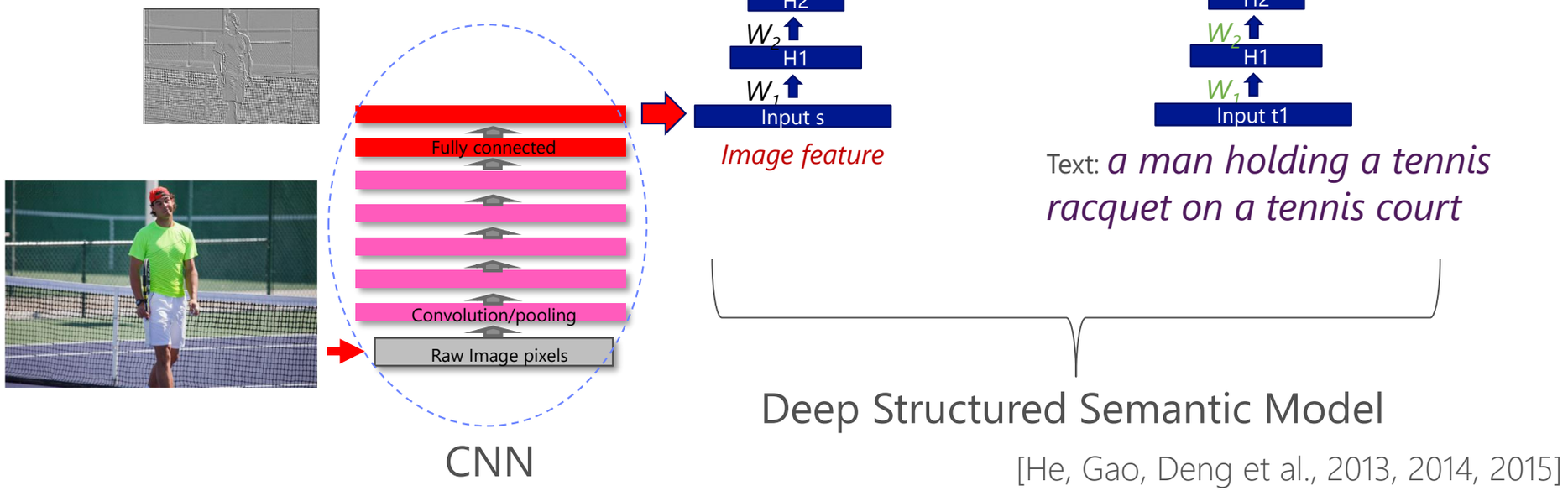
500. a room with stove on the floor

[Fang, et al., CVPR 2015]

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]

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

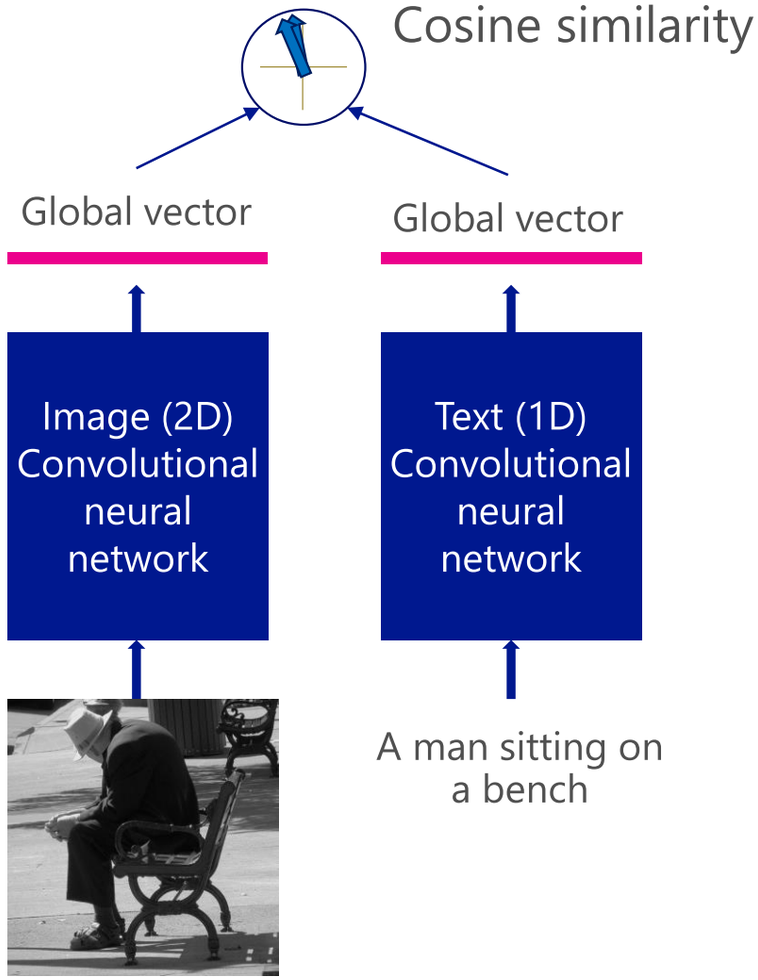
Caption probability: $P(D|Q) = \frac{\exp(\gamma R(Q, D))}{\Sigma_{D' \in \mathbb{D}} \exp(\gamma R(Q, D'))}$

Candidate captions ↗ ↖ Smoothing factor

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

Correct caption

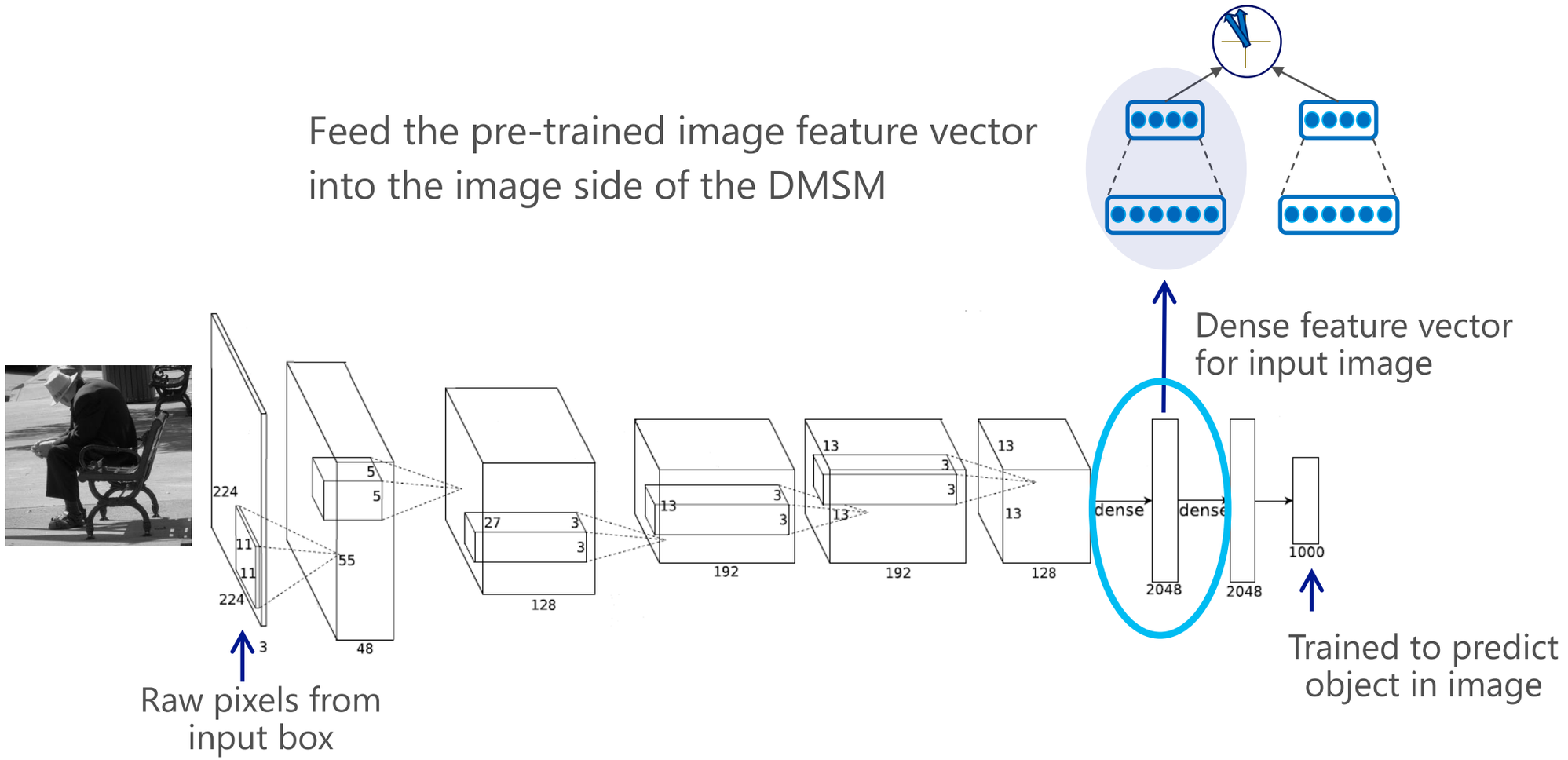
Huang, He, Gao, Deng, Acero, Heck, "Learning Deep Structured Semantic Model for Web Search," CIKM, 2013



Serves as a semantic matching checker.

The convolutional network at the image side

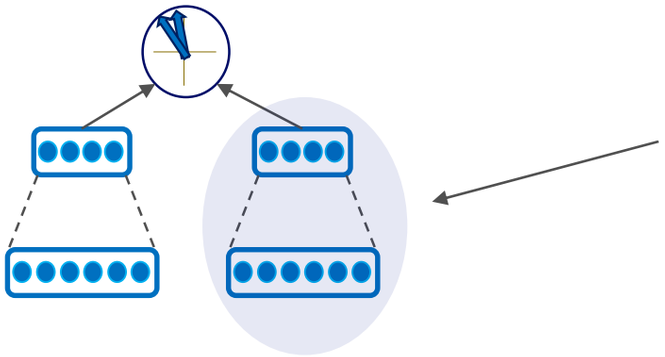
Feed the pre-trained image feature vector into the image side of the DMSM



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

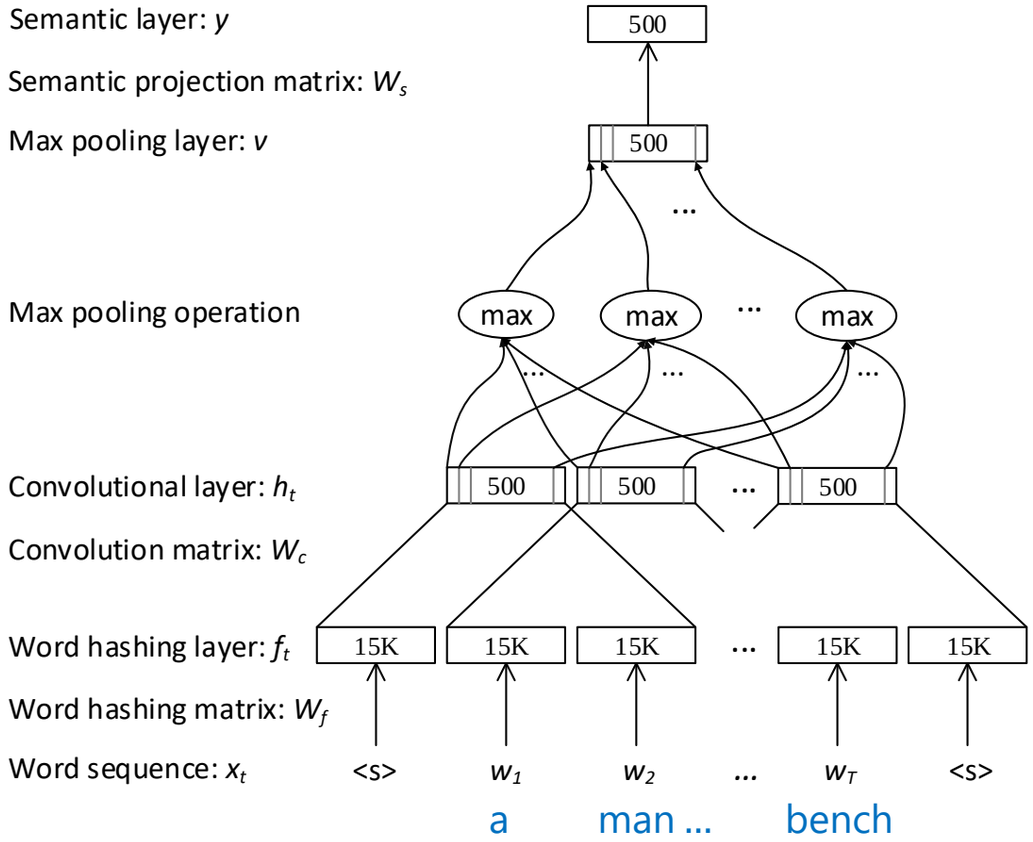


Figure Credit: [Shen, He, Gao, Deng, Mesnil, WWW, April 2014]

COCO Captioning Challenge 2015

Human judgment is the ultimate metric

e.g., Turing Test

Microsoft and Google shared the 1st prize.

Retrieval-based method

human generated caption

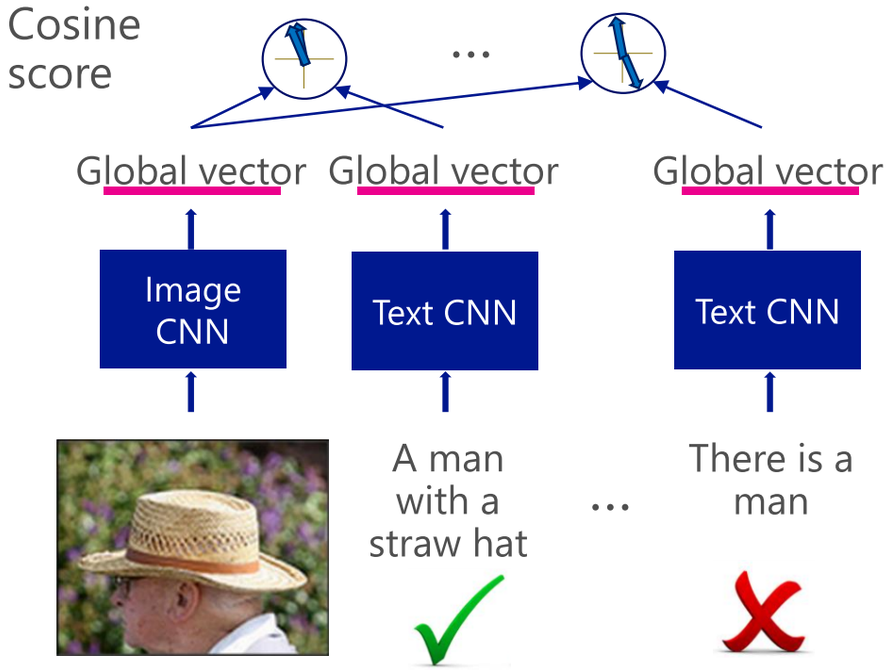
Microsoft COCO
Common Objects in Context

	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%**

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



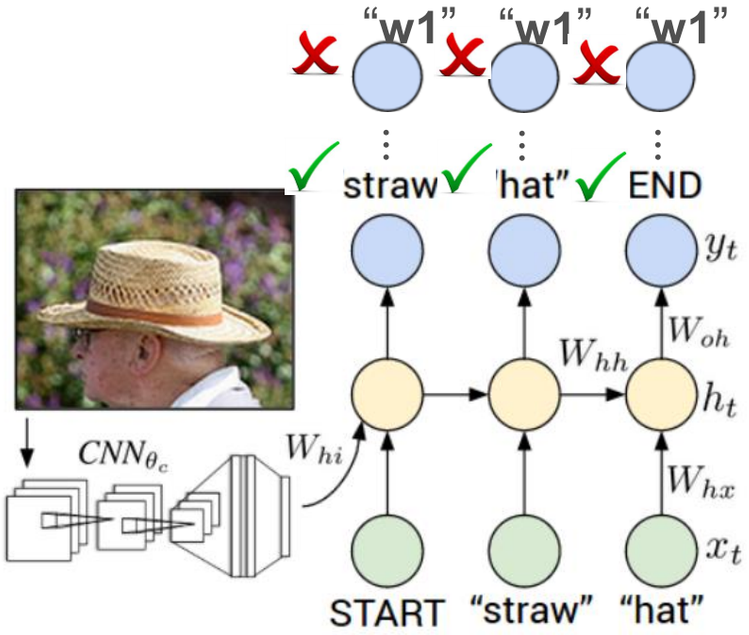
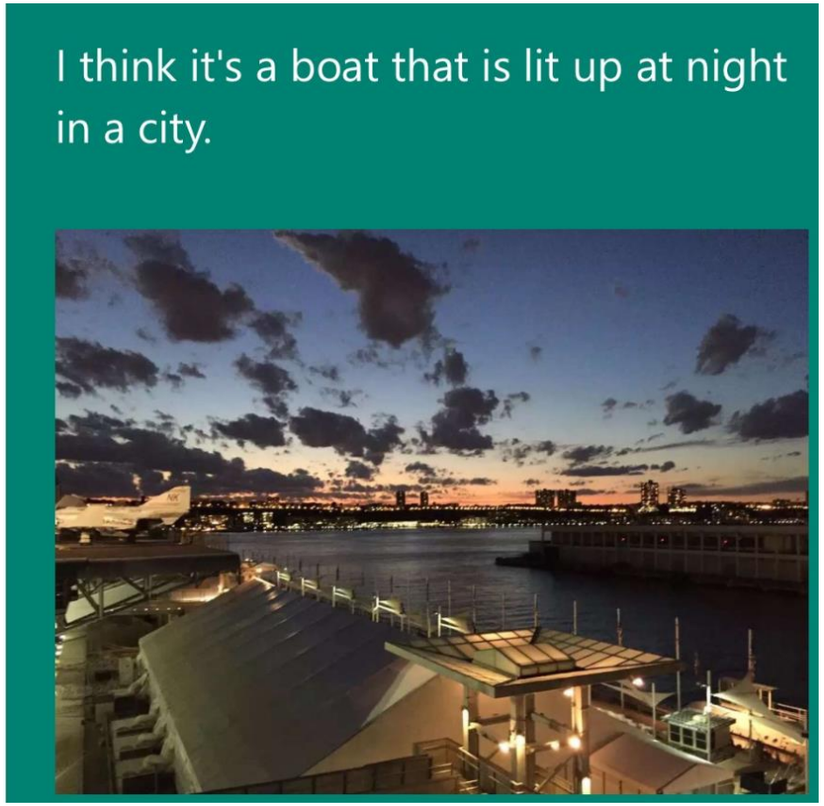


Image Credit: Karpathy and Fei-Fei 2015

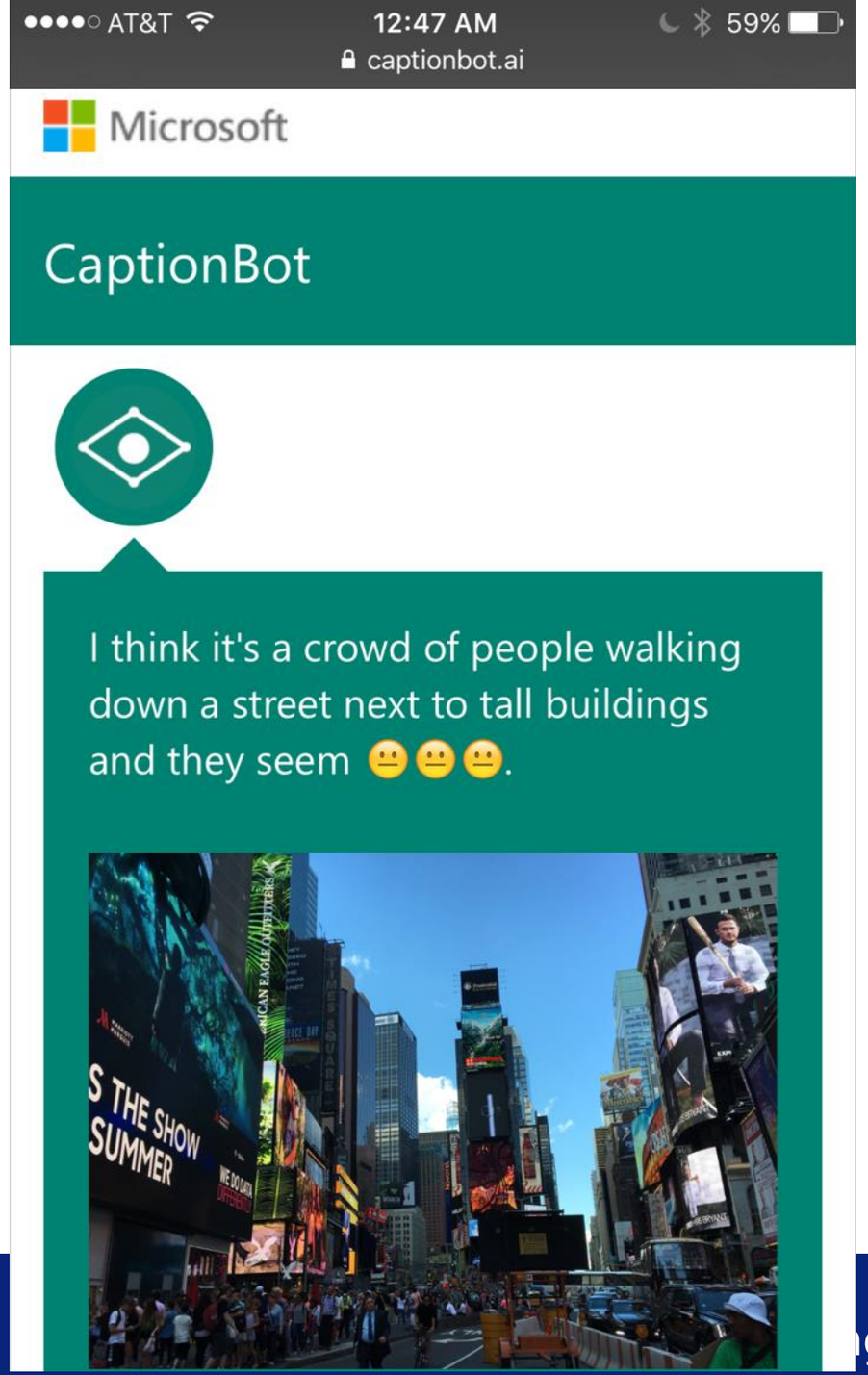
DMSM focuses on semantics rather than syntax. E.g., ensures the reference (*semantically interesting*) scores higher than generic ones (grammatically correct but *semantically incorrect or boring*), while MRNN focus on syntax more.

Public App: CaptionBot

http://CaptionBot.ai



[Tran, He, Zhang, Sun, Carapcea, Thrasher, Buehler, Sienkiewicz, "Rich Image Captioning in the Wild," DeepVision, CVPR 2016]



More Examples from CaptionBot

CaptionBot

I think it's a large body of water with a city in the background.

I think it's a man riding a horse jumping over an obstacle.

More Examples

I think it's a man flying a kite on the beach.

I think it's a colorful bird perched on a tree branch.

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

From Captioning to Question Answering

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

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

Image Question Answering (IQA)

Answer:
dogs

Caption vs. QA: need reasoning

Image QA: reasoning is the key.

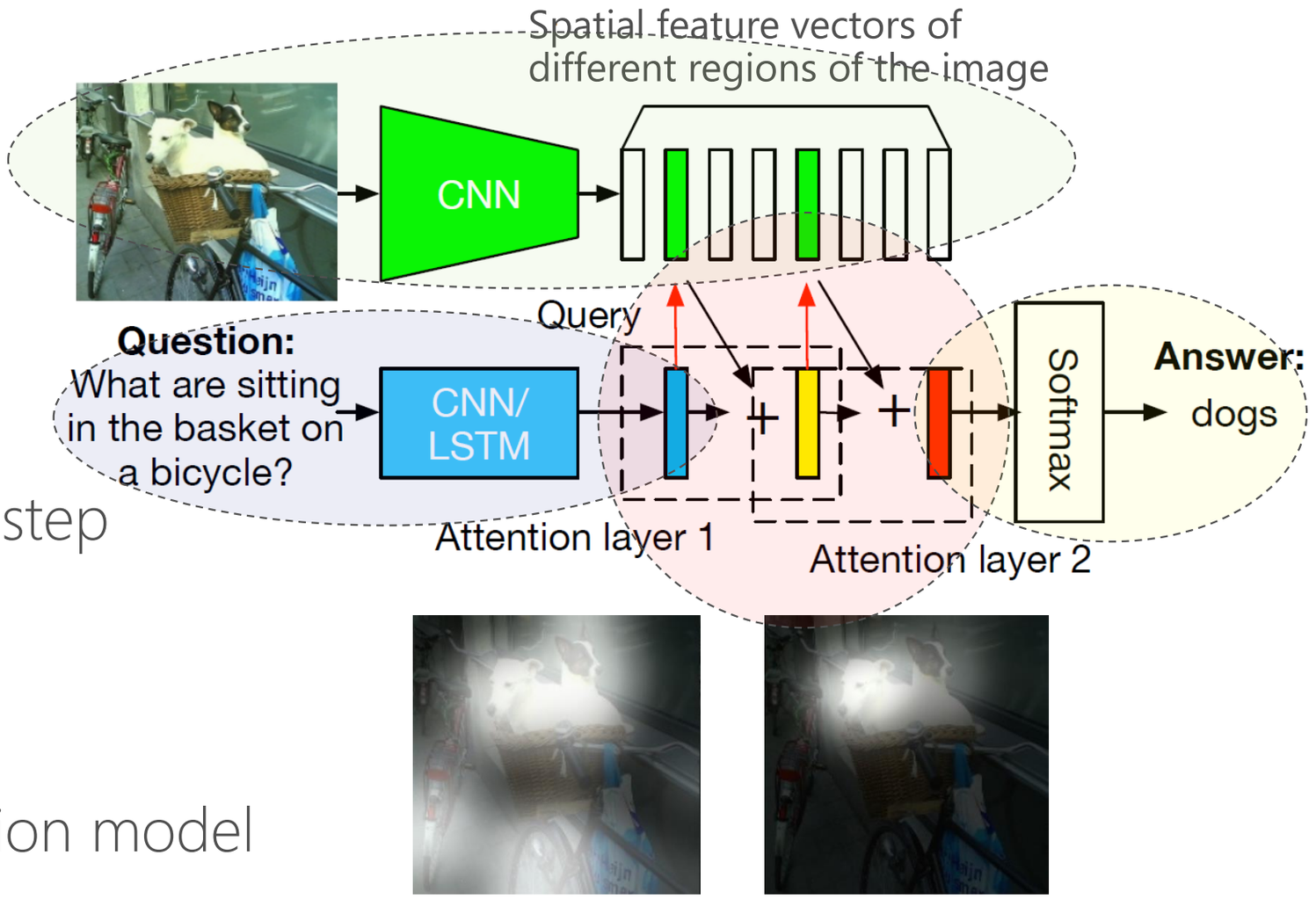
Question:
What are sitting in the basket on a bicycle?

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

Answer:
dogs

Stacked Attention 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

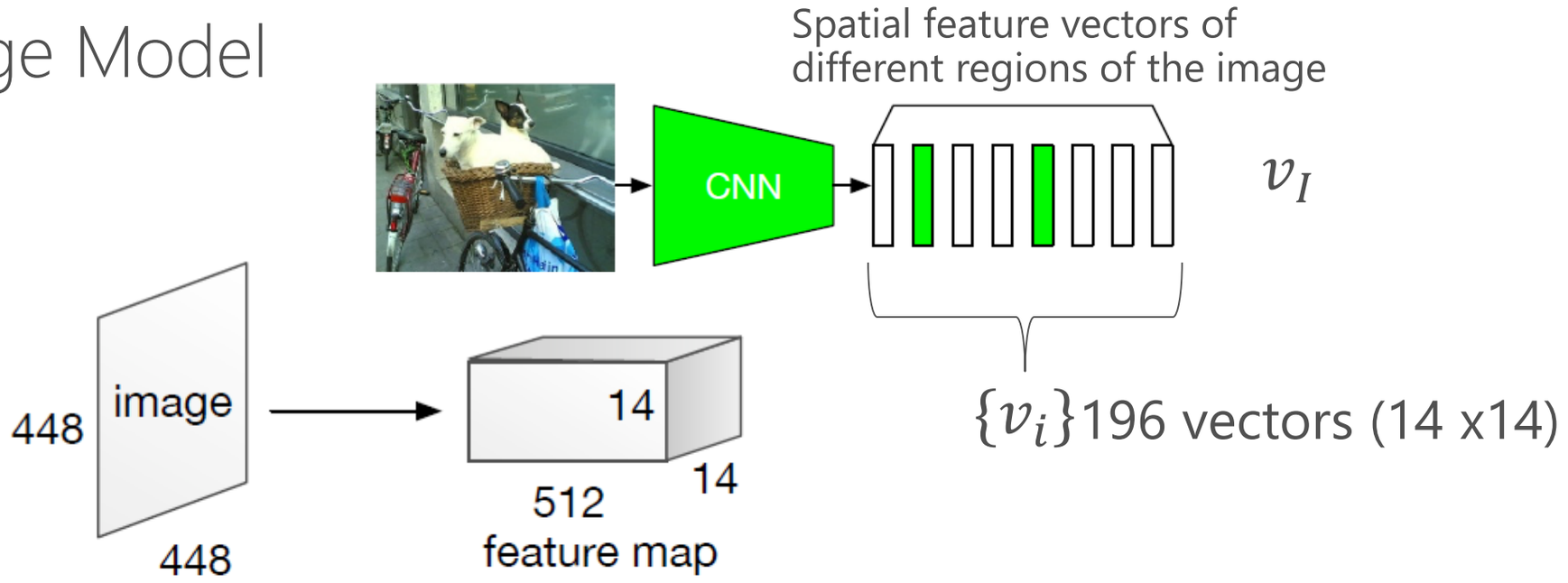


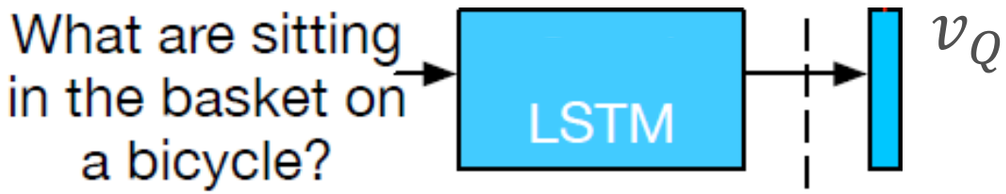
Figure 2: CNN based image model

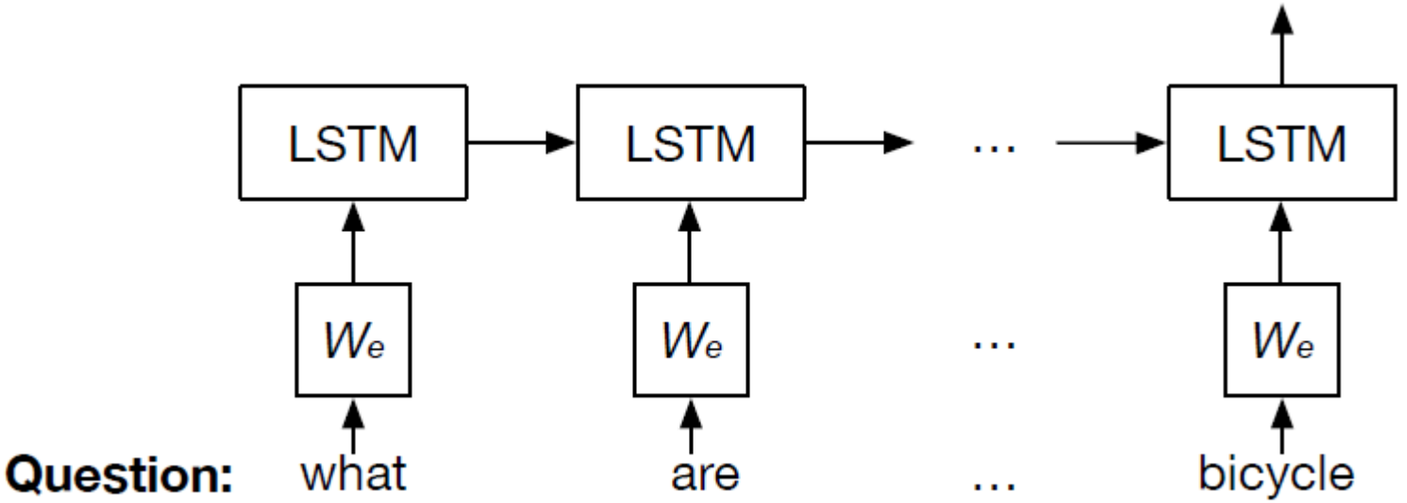
$$f_I = \text{CNN}_{vgg}(I). \qquad v_I = \tanh(W_I f_I + b_I)$$

2. The question model in the SAN

- Question Model

  Code the question into a vector using a LSTM

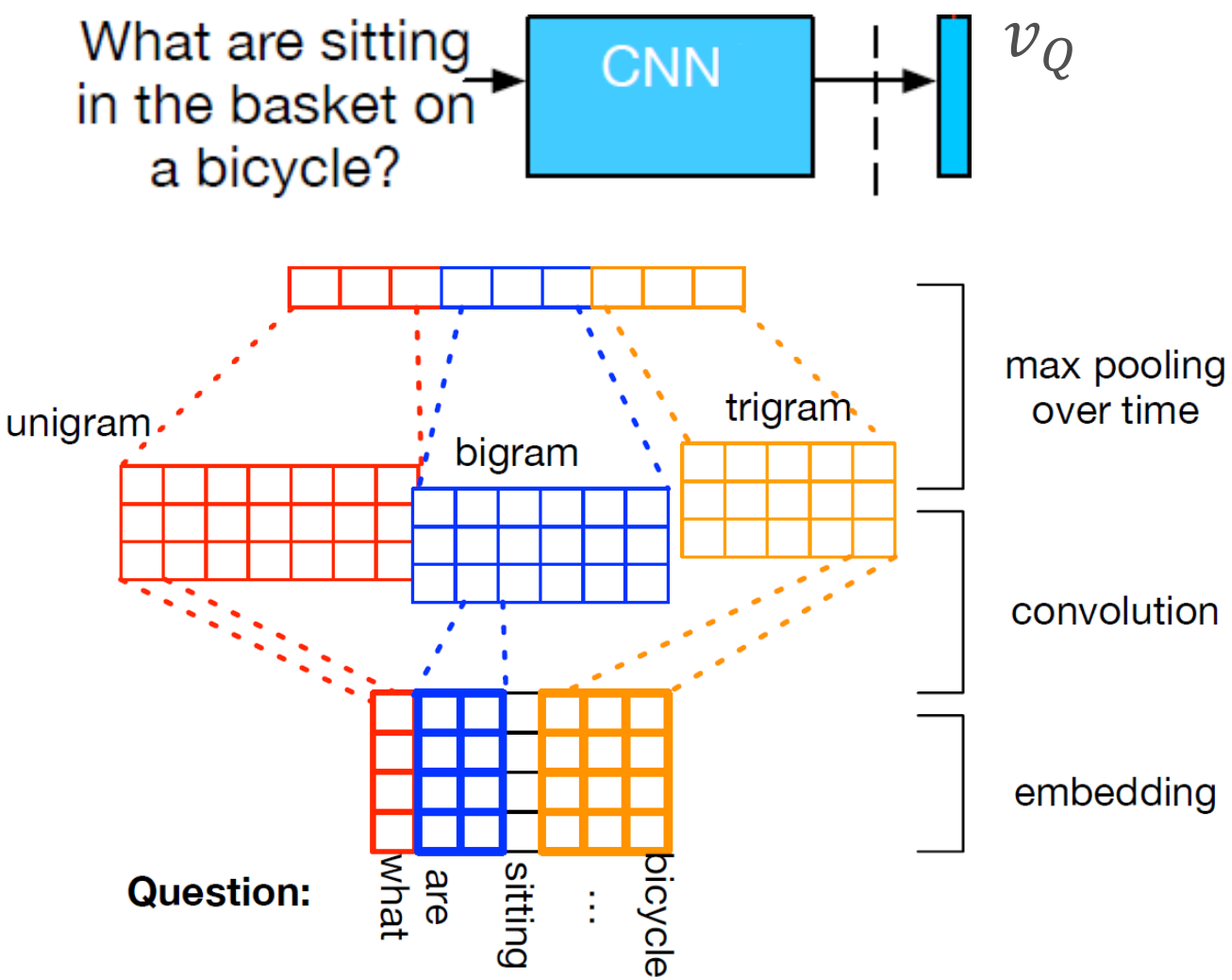




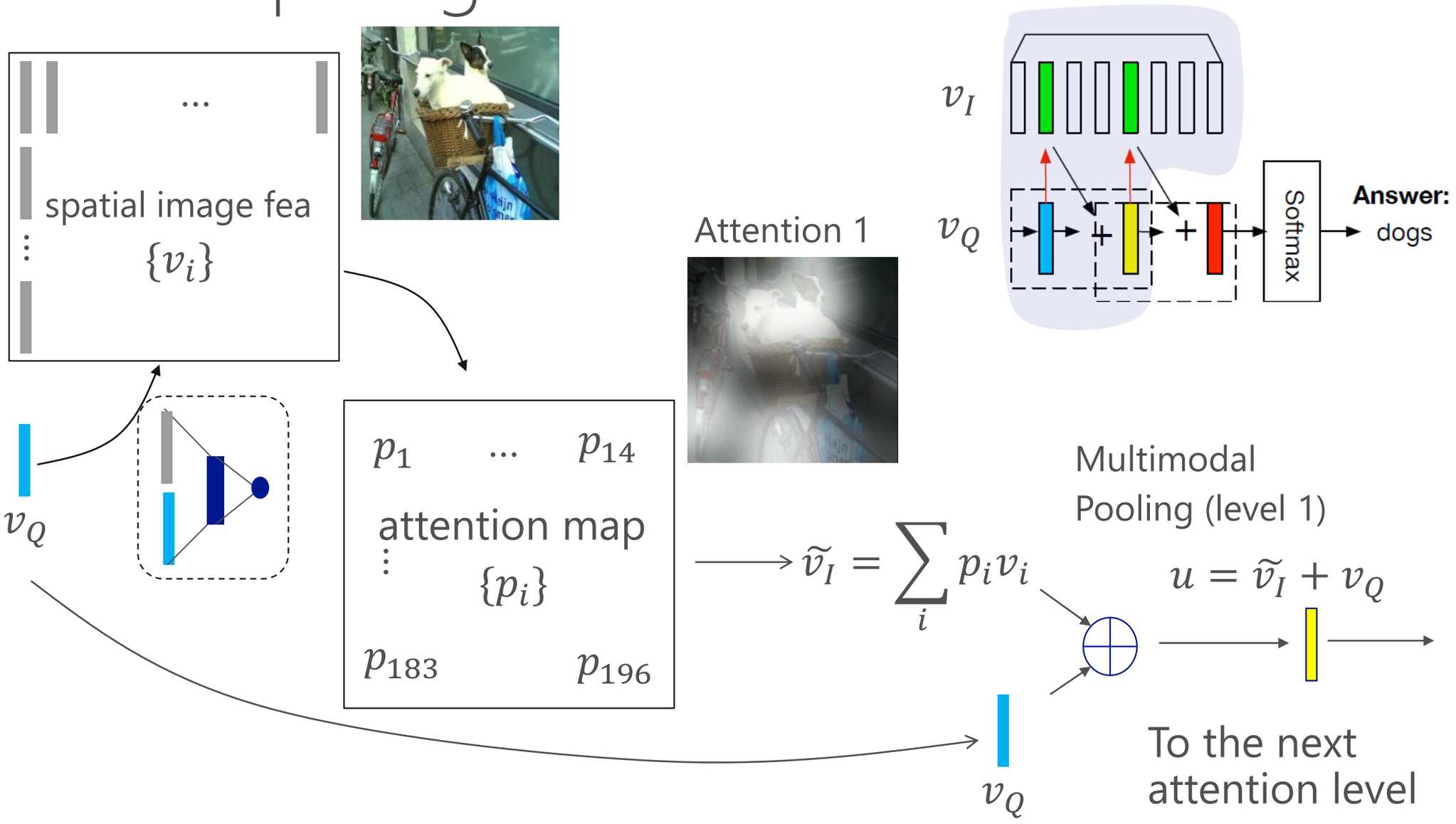
2. The question model in the SAN (alternative)

- Question Model

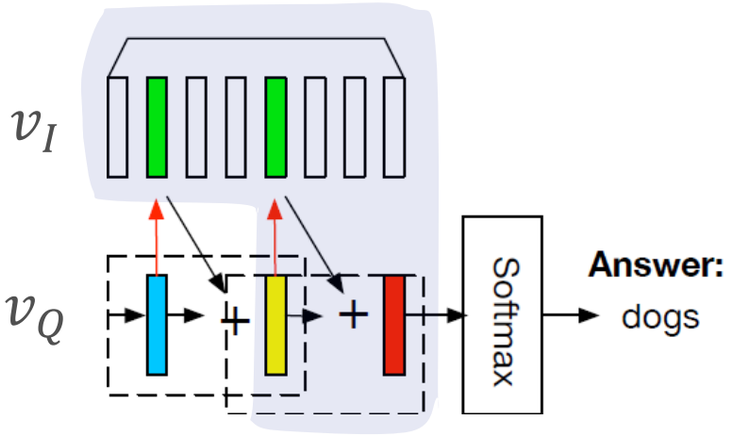
  Code the question into a vector using a CNN



3. SAN: Computing the 1st level attention



3. SAN: Compute the 2nd level attention



Attention 2

spatial img fea $\{v_i\}$

u

p_1 ... p_{14}

attention map $\{p_i\}$

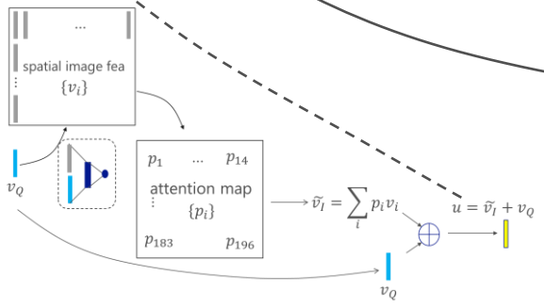
p_{183} p_{196}

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

Multimodal Pooling (level 2)

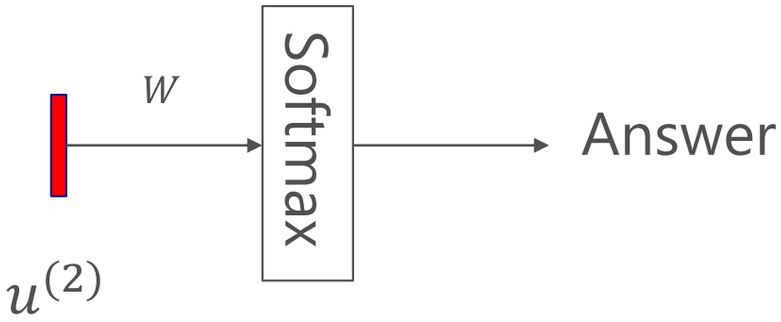
$$u^{(2)} = \tilde{v}_I^{(2)} + u$$

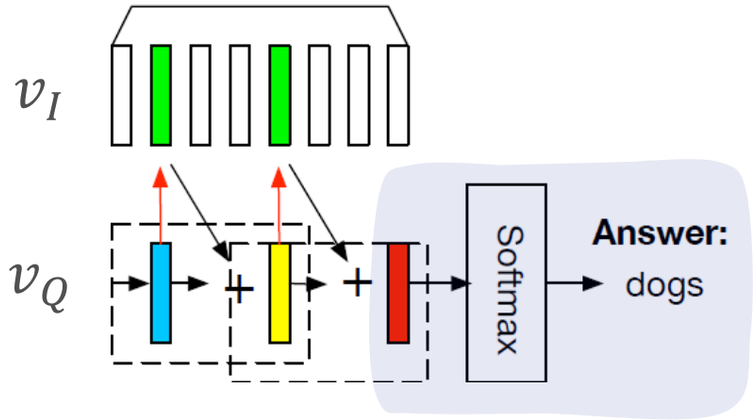
To the answer predictor



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**

Other relevant work

- Andreas, Rohrbach, Darrell, Klein, "Neural Module Networks,", CVPR 2016
- Noh, Seo, Han, "Image Question Answering Using Convolutional Neural Network With Dynamic Parameter Prediction," CVPR 2016
- Shih, Singh, Hoiem, "Where to Look: Focus Regions for Visual Question Answering," CVPR 2016
- Wu, Wang, Shen, Dick, Hengel, "Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge From External Sources," CVPR 2016
- Tapaswi, Zhu, Stiefelhagen, Torralba, Urtasun, Fidler, "MovieQA: Understanding Stories in Movies Through Question-Answering," CVPR 2016
- Zhu, Groth, Bernstein, Fei-Fei, "Visual7W: Grounded Question Answering in Images," CVPR 2016
- Kafle, Kanan, "Answer-Type Prediction for Visual Question Answering," CVPR 2016

And more work in the Visual Question Answering (VQA) Challenge Workshop, CVPR2016