语言与视觉多模态智能的进展

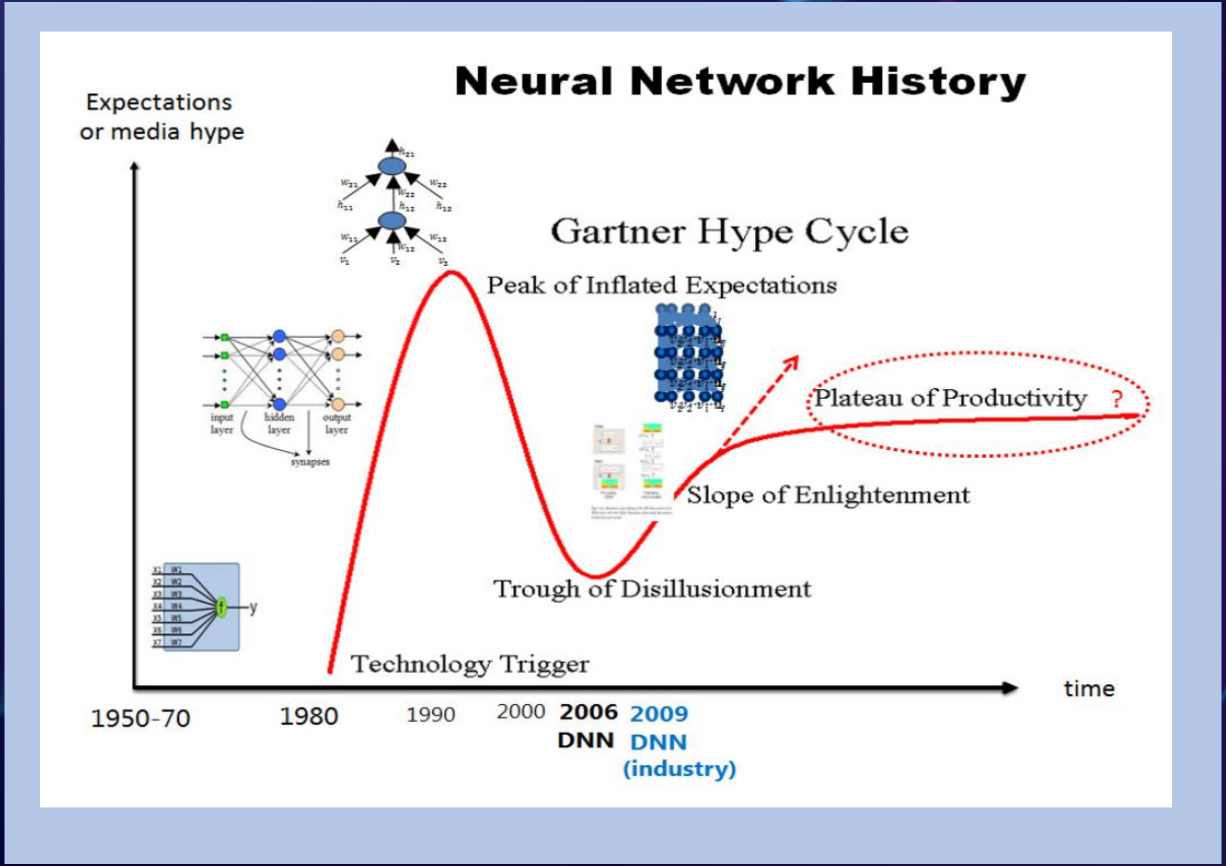
何晓冬

IEEE/CAAI Fellow
京东人工智能研究院常务副院长
深度学习和语音及语言实验室主任

提纲

- 近来的一些进展
  - 深度学习
  - 语音和自然语言理解
- 语言+视觉多模态智能
  - 图像描述（Image-to-text Captioning）
  - 视觉问答（Visual Question Answering）
  - 基于文字描述合成图像（Text-to-image Synthesis）

深度学习的发展



Geoff
Hinton

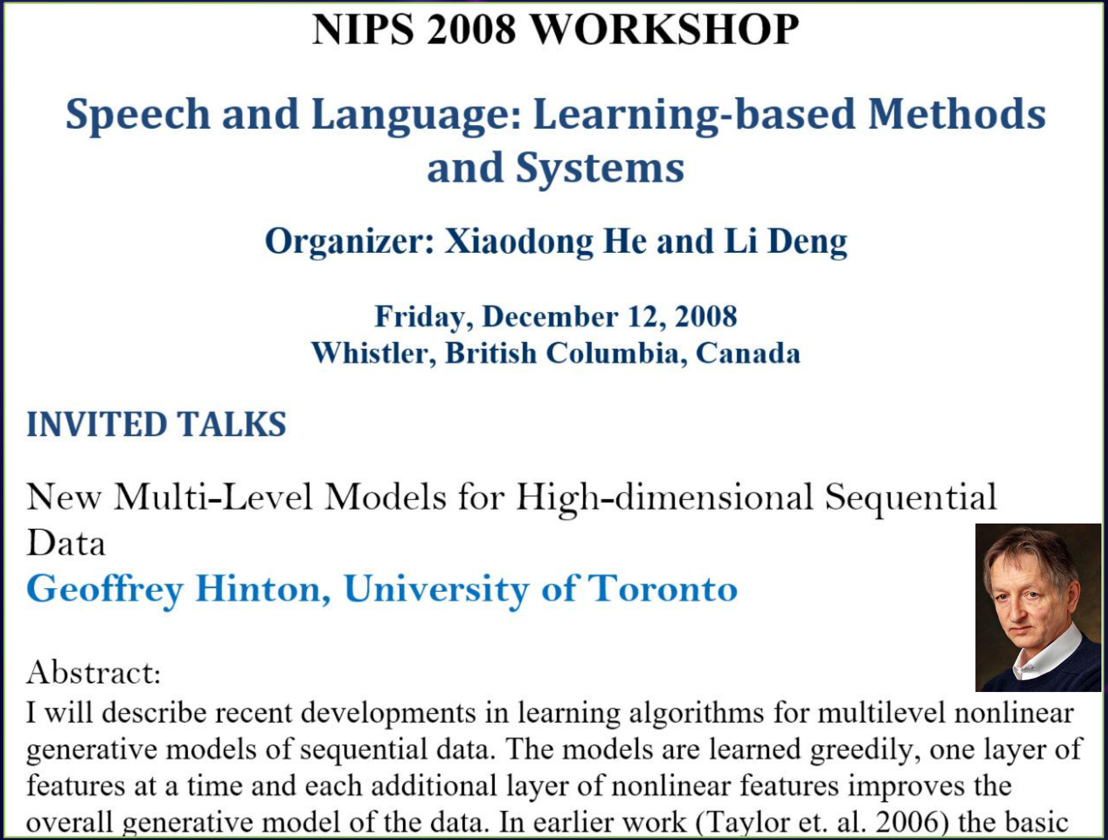
Yoshua
Bengio

Yann
LeCun

2018

从学术界到工业界：Scale up!

NIPS08 workshop 邀请Hinton来做报告并探讨与工业界合作



NIPS 2008 WORKSHOP

Speech and Language: Learning-based Methods and Systems

Organizer: Xiaodong He and Li Deng

Friday, December 12, 2008
Whistler, British Columbia, Canada

INVITED TALKS

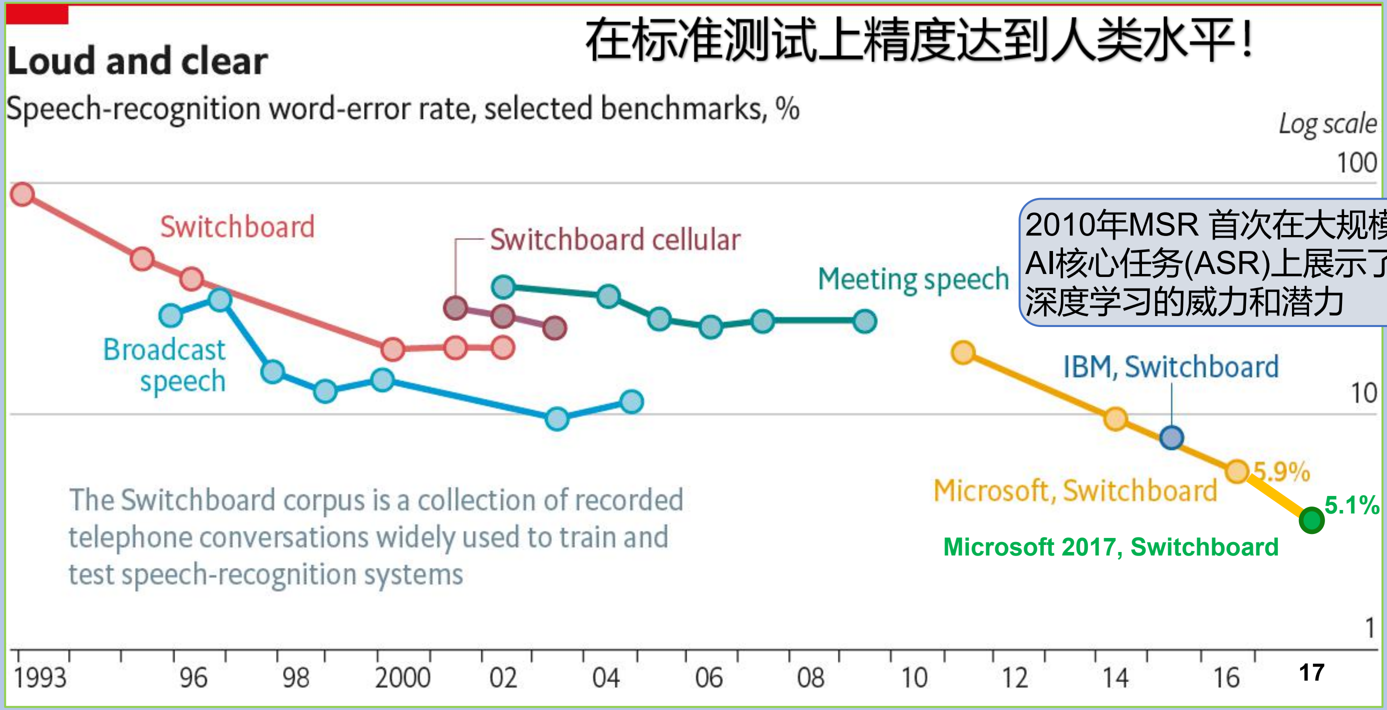
New Multi-Level Models for High-dimensional Sequential Data

Geoffrey Hinton, University of Toronto

Abstract:
I will describe recent developments in learning algorithms for multilevel nonlinear generative models of sequential data. The models are learned greedily, one layer of features at a time and each additional layer of nonlinear features improves the overall generative model of the data. In earlier work (Taylor et. al. 2006) the basic

2018

语音识别

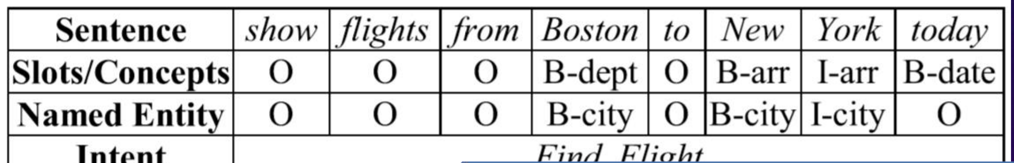


2010年MSR 首次在大规模AI核心任务(ASR)上展示了深度学习的威力和潜力

2018

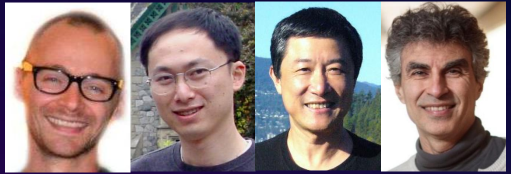
语言理解/语义槽值提取

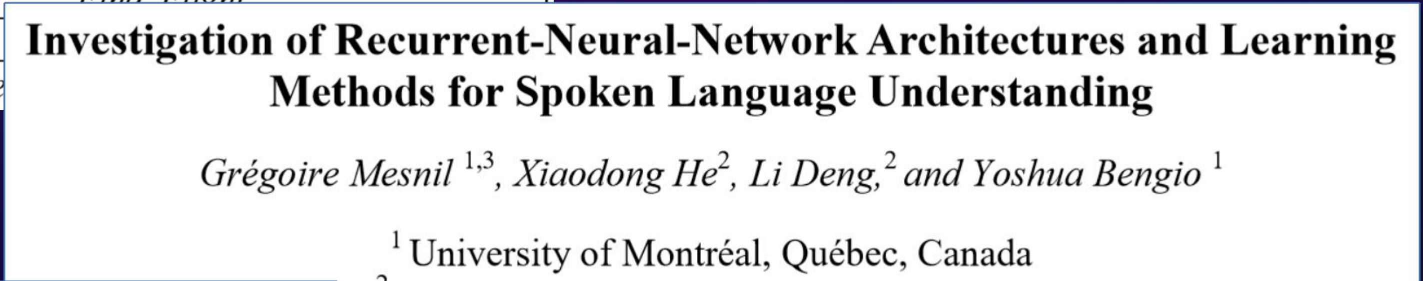
2013年成功应用RNN for SLU



Sentence	*show*	*flights*	*from*	*Boston*	*to*	*New*	*York*	*today*
Slots/Concepts	O	O	O	B-dept	O	B-arr	I-arr	B-date
Named Entity	O	O	O	B-city	O	B-city	I-city	O
Intent	*Find_Flight*							
Domain								

Table 1. ATIS utterance e





Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding

Grégoire Mesnil [1,3], *Xiaodong He*[2], *Li Deng,*[2] *and Yoshua Bengio* [1]

[1] University of Montréal, Québec, Canada

[2] Microsoft Research, Redmond, WA, USA

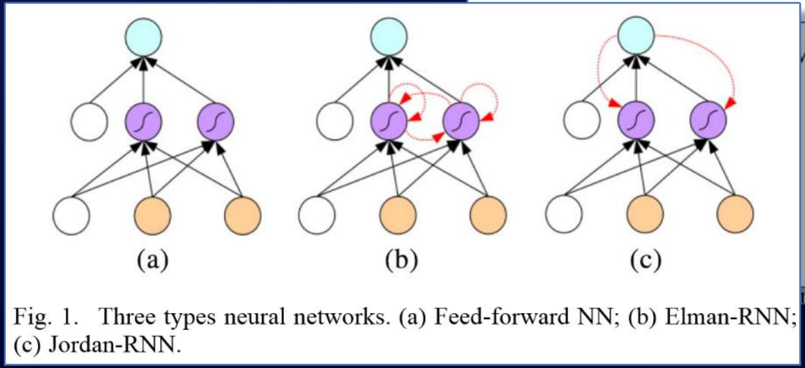
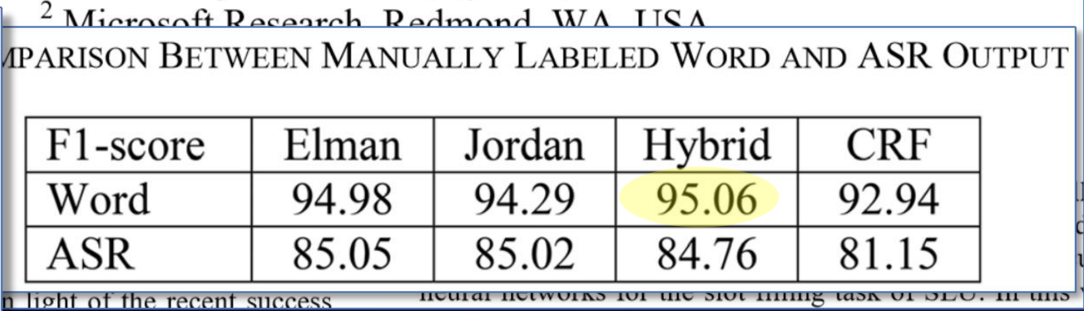


Fig. 1. Three types neural networks. (a) Feed-forward NN; (b) Elman-RNN; (c) Jordan-RNN.



MPARISON BETWEEN MANUALLY LABELED WORD AND ASR OUTPUT

F1-score	Elman	Jordan	Hybrid	CRF
Word	94.98	94.29	95.06	92.94
ASR	85.05	85.02	84.76	81.15

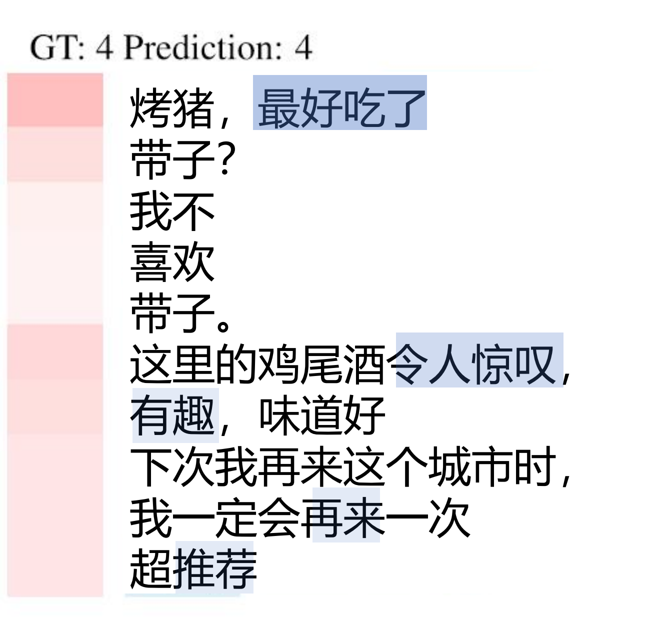
n light of the recent success   neural networks for the slot filling task of SLU. In this

[Mesnil, He, Deng, Bengio, InterSpeech2013]

语言理解/语义意图分类

Hierarchical Attention Net (HAN)

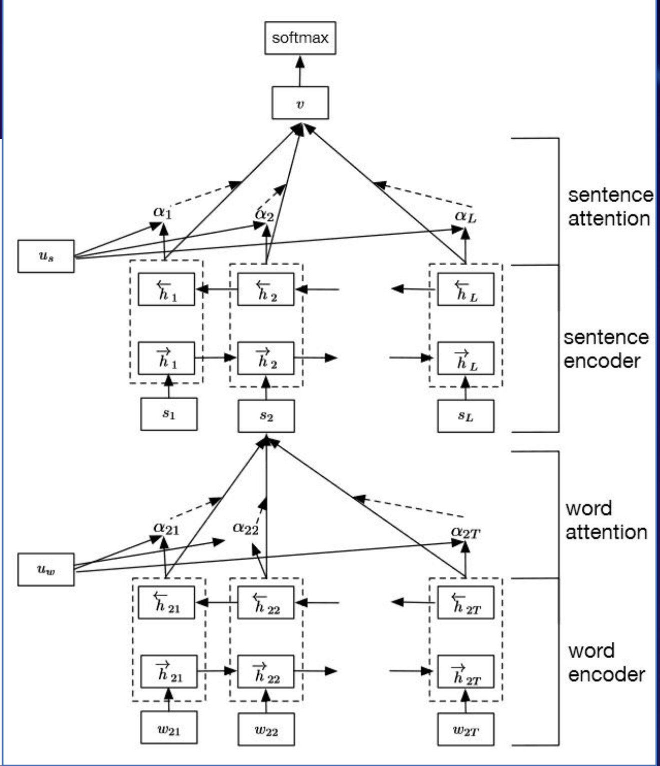
我们于2016年提出的层次化注意力模型 (HAN)能在词、句子、段落、等多个层面来建模理解语言，判断意图，并通过对神经元激活的可视化来给出一定程度的可解释性。



GT: 4 Prediction: 4

烤猪，最好吃了
带子?
我不
喜欢
带子。
这里的鸡尾酒令人惊叹,
有趣，味道好
下次我再来这个城市时,
我一定会再来一次
超推荐

Introducing self-attention



2018

[Yang, Yang, Dyer, He, Smola, Hovy, “*HAN*”, NAACL2016] (citation: 1200)

语言理解/语义的表征

从自然语言中抽取出语义并将其投影到语义空间以帮助搜索、推荐、分类、问答等应用

抽象的语义表征

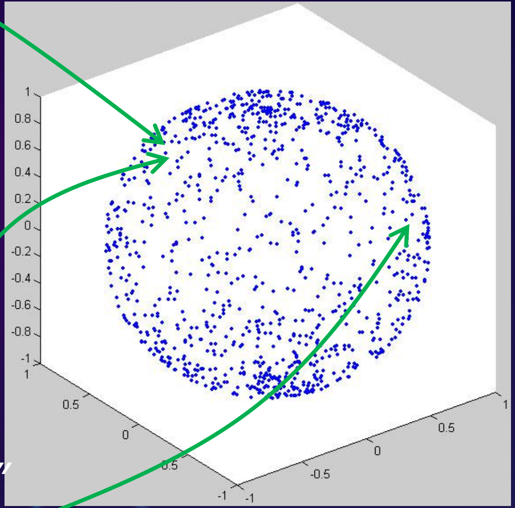
通过深度神经网络逐步抽取
语义上的不变性 (invariance)

神经网络

输入

自然语言的描述 *“小明快递了一袋苹果给外公”*

语义空间



语义相似的描述 *“外公从小明那收到了袋红富士”*

语义不同的描述 *“小明送给女友最新一代的苹果 X”*

[Huang, He, Gao, Deng, Acero, Heck, “*DSSM*”, CIKM2013]

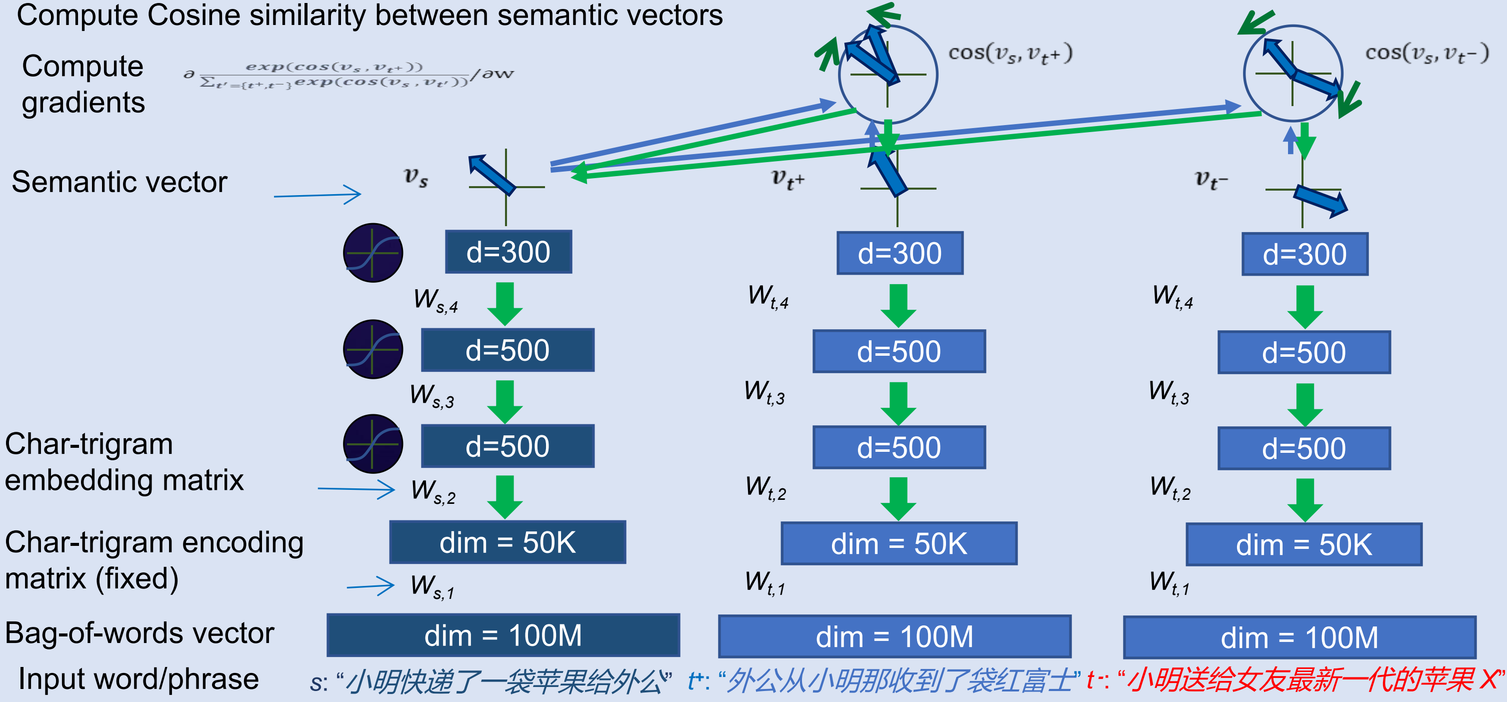
2018

DSSM: 深度结构化语义模型

基于相对相似度的训练目标函数:

Compute Cosine similarity between semantic vectors

Compute gradients $\partial \frac{exp(cos(v_s, v_{t^+}))}{\sum_{t'=\{t^+,t^-\}} exp(cos(v_s, v_{t'}))} / \partial w$



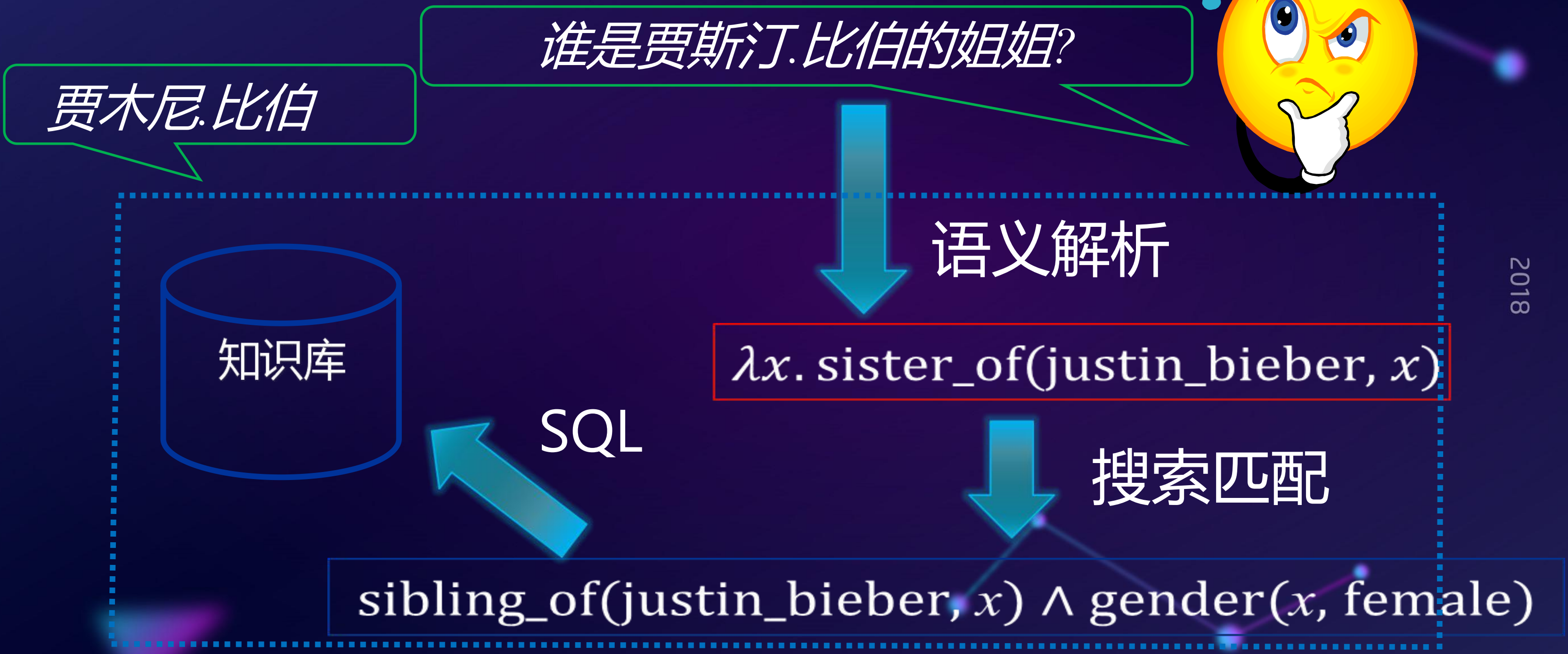
s: "小明快递了一袋苹果给外公" t^+: "外公从小明那收到了袋红富士" t^-: "小明送给女友最新一代的苹果 X"

[Huang, He, Gao, Deng, Acero, Heck, "*DSSM*", CIKM2013;
Shen, He, Gao, Deng, Mesnil, "*CDSSM*", WWW2014&CIKM2014】

2018

知识推理及问答

在连续向量空间表达知识、解析语义、执行推理和应答



谁是贾斯汀.比伯的姐姐?

贾木尼.比伯

语义解析

$\lambda x.\,\mathrm{sister\_of}(\mathrm{justin\_bieber}, x)$

搜索匹配

$\mathrm{sibling\_of}(\mathrm{justin\_bieber}, x) \wedge \mathrm{gender}(x, \mathrm{female})$

SQL

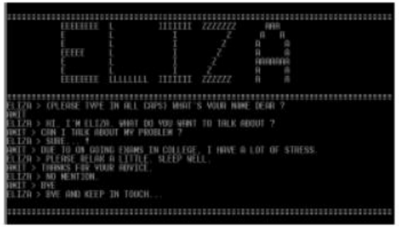
知识库

2018

[Yih, He, Meek, ACL2014; Yih, Chang, He, Gao, ACL2015; Golub & He, EMNLP2016;…]

对话机器人

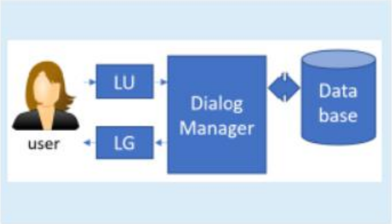
Early Chatbots



```
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE...!
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...
```

1966

Task-Completion



1990's

Personal Assistant

2012

Social Chatbots

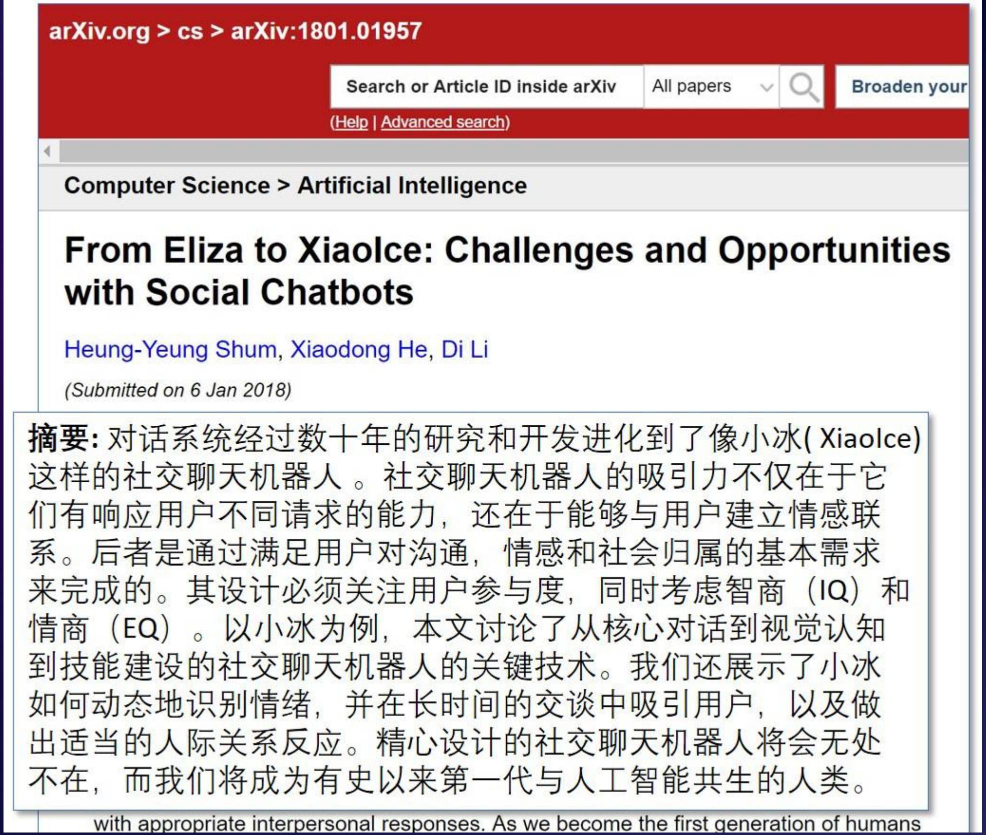
2014

From mimicking humans' behavior,
to understanding humans' requests,
to serve humans' proactive and reactive needs,
and to building emotional connection with humans

从语言理解、问答、到人机对话进展显著

“ 对话机器人不仅需要响应用户的请求，完成任务，还需要满足用户对沟通和情感的需求，与用户建立情感联系。” “我 们将成为有史以来第一代与 AI 共生的人类。”

— *“从Eliza到小冰：社交对话机器人的机遇和挑战”*，Harry Shum, Xiaodong He, Di Li, in FITEE 2018



arXiv.org > cs > arXiv:1801.01957

Computer Science > Artificial Intelligence

From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots

Heung-Yeung Shum, Xiaodong He, Di Li

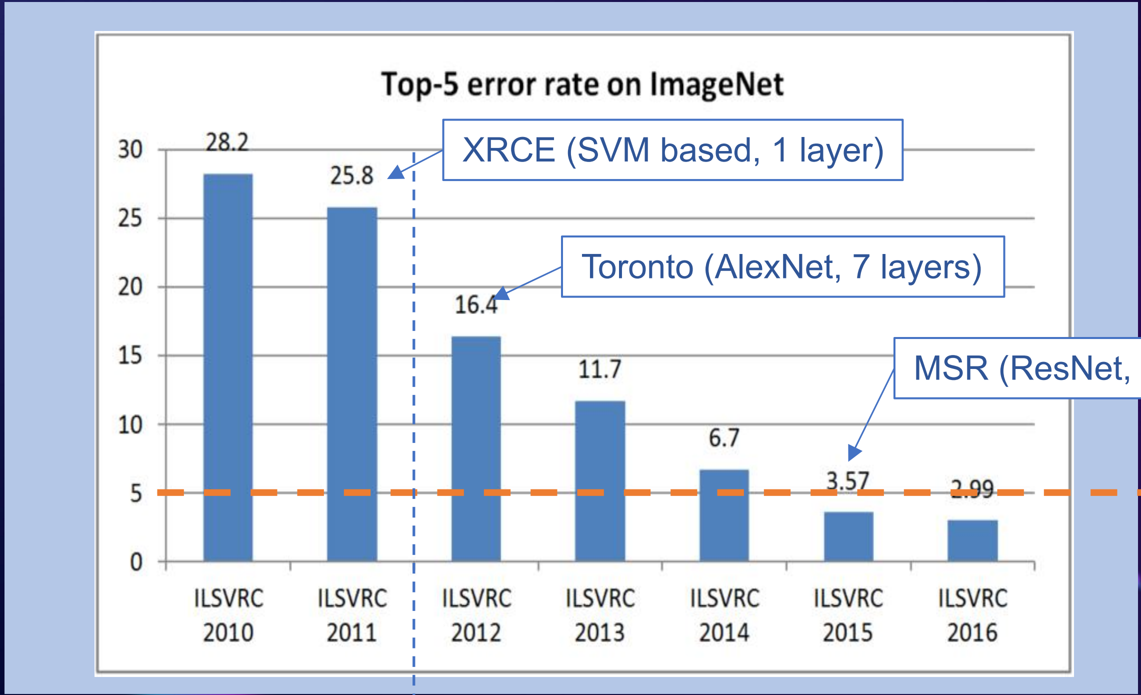
(Submitted on 6 Jan 2018)

摘要: 对话系统经过数十年的研究和开发进化到了像小冰(XiaoIce)这样的社交聊天机器人 。社交聊天机器人的吸引力不仅在于它们有响应用户不同请求的能力，还在于能够与用户建立情感联系。后者是通过满足用户对沟通，情感和社会归属的基本需求来完成的。其设计必须关注用户参与度，同时考虑智商（IQ）和情商（EQ） 。以小冰为例，本文讨论了从核心对话到视觉认知到技能建设的社交聊天机器人的关键技术。我们还展示了小冰如何动态地识别情绪，并在长时间的交谈中吸引用户，以及做出适当的人际关系反应。精心设计的社交聊天机器人将会无处不在，而我们将成为有史以来第一代与人工智能共生的人类。

with appropriate interpersonal responses. As we become the first generation of humans

视觉智能

AI图像识别在ImageNet测试上达到人类水平



XRCE (SVM based, 1 layer)

Toronto (AlexNet, 7 layers)

MSR (ResNet, 152 layers)

人类物体识别错误率约5%

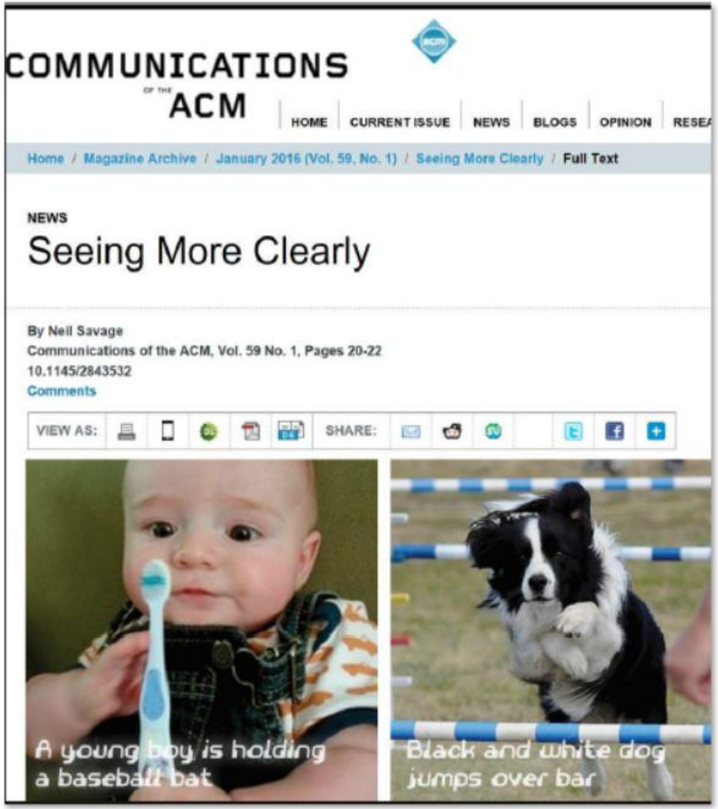
2012 前，大都是线性模型

2012后，主流模型是深度神经网络

[Fei-Fei Li +]IMAGENET

(1000 类物体识别测试)

语言+视觉



COMMUNICATIONS OF THE ACM

HOME
CURRENT ISSUE
NEWS
BLOGS
OPINION
RESEA

Home / Magazine Archive / January 2016 (Vol. 59, No. 1) / Seeing More Clearly / Full Text

NEWS

Seeing More Clearly

By Neil Savage
Communications of the ACM, Vol. 59 No. 1, Pages 20-22
10.1145/2843532
Comments

VIEW AS:
SHARE:
A young boy is holding a baseball bat
Black and white dog jumps over bar

“With careful training, these things (object recognition) actually work very well,” – *Rob Fergus*

“The complete level, on par with an adult, I think is going to be a long way off,” – *Fei-Fei Li*

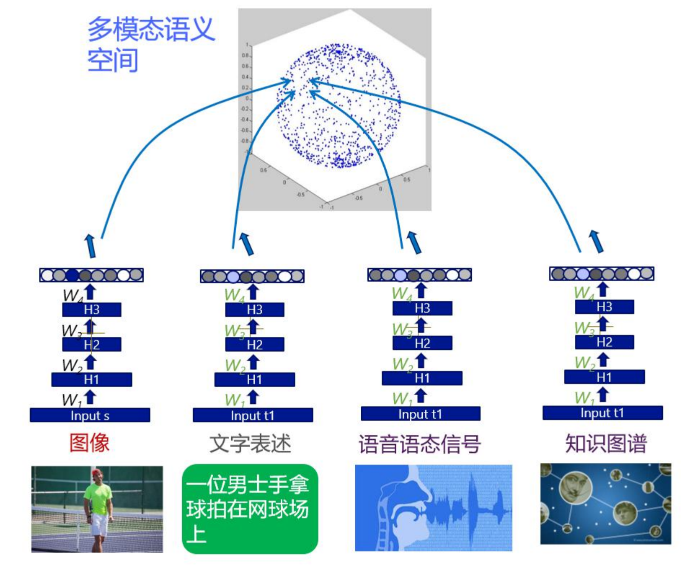
“The overall picture should have the same semantic value as the description,” – *Xiaodong He*

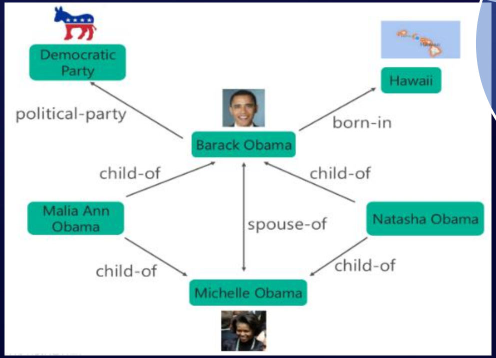
“If you really understood the image, you could answer a question about it.” - *Richard Zemel*

CACM January 2016 (Vol. 59,No. 1)

多模态智能: 语音,语言,视觉,知识 +

贝拉克·侯赛因·奥巴马，美国民主党籍政治家，第44任美国总统，为美国历史上第一位非裔美国人（美国黑人）总统。





语言-视觉多模态：三个研究视角

- 表征预训练
  - 从原始的语言或图像信号中提取语义表征。往往是将语言和图像信号通过预训练映射到一个连续向量空间。

- 跨模态表征融合与印证
  - 在连续向量空间融合多个模态信息。往往是通过跨模态池化模型来融合语言和图像的表征（multimodal pooling），及通过注意力模型来为语言和图像的结构建立联接与印证（grounding）

- 基于多模态任务的模型优化
  - 基于特定任务设计优化目标函数及优化算法。比如包括cross-entropy，BLEU；Reinforcement learning，GAN

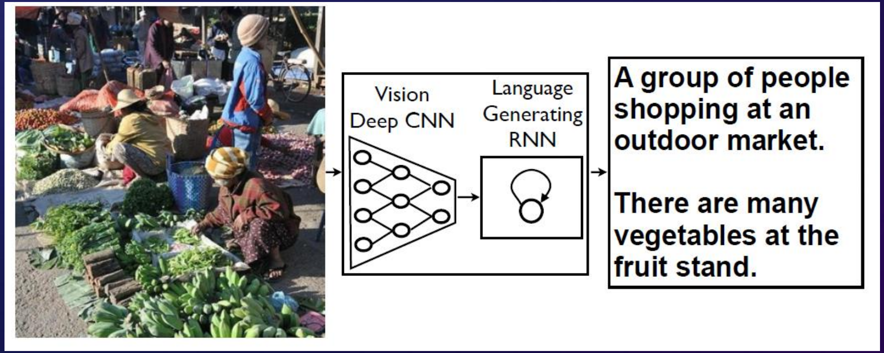
语言-视觉多模态任务

- 图像到文本描述（image-to-text / image captioning）
  - 理解图像的内容，生成自然语言来描述图像内容
- 视觉-文本问答 （visual question answering）
  - 基于对图像的理解回答相关的文本问题
- 文本到图像生成（text-to-image synthesis）
  - 基于对文字描述的理解以生成相应的图像
- 语言-视觉导航，视觉对话，跨模态信息检索 …

2018

图像描述（Image Captioning）

自CVPR 2015 以来的一系列工作



A group of people shopping at an outdoor market.

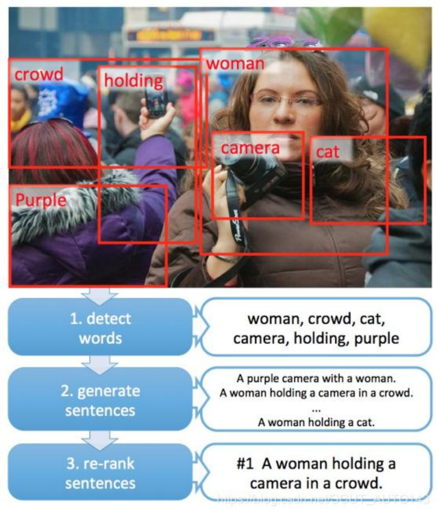
There are many vegetables at the fruit stand.

End-to-End Paradigm：
1. Encode image to vector representations （学习图像表征）
2. generate sentences（生成句子）

[Vinyals, Toshev, Bengio, Erhan，“Show and Tell: A Neural Image Caption Generator,” CVPR2015]

Cascade Paradigm：
1.detect words（检测关键物体、概念）
2.generate sentences（生成候选句子）
3.Semantic re-rank sentences（按语义表征排序）

[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, et al.,“From Captions to Visual Concepts and Back,” CVPR2015]



1. detect words — woman, crowd, cat, camera, holding, purple

2. generate sentences — A purple camera with a woman.
A woman holding a camera in a crowd.
...
A woman holding a cat.

3. re-rank sentences — #1 A woman holding a camera in a crowd.

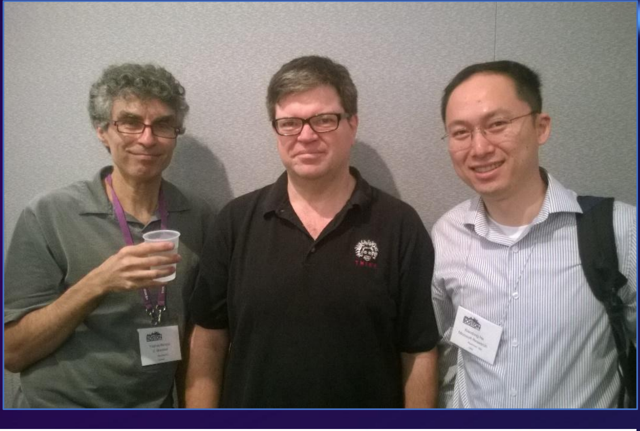
DeepVision

DeepVision: Deep Learning in Computer Vision, 2nd edition

CVPR 2015 Workshop
June 11th, 2015, Boston, USA

Organizing Committee
Y. LeCun, Facebook AI&NYU
J. Alvarez, NICTA, Australia
Y. Li, NICTA, Australia
F. Porikli, ANU/NICTA, Australia

Invited Speakers
Yoshua Bengio, U. Montreal
Xiaodong He, MS Research
Stephen Jones, NVIDIA
Randall C. O'Reilly, U. Colorado
Rahul Sukthankar, Google
Andrea Vedaldi, Oxford
Xiaogang Wang, CUHK

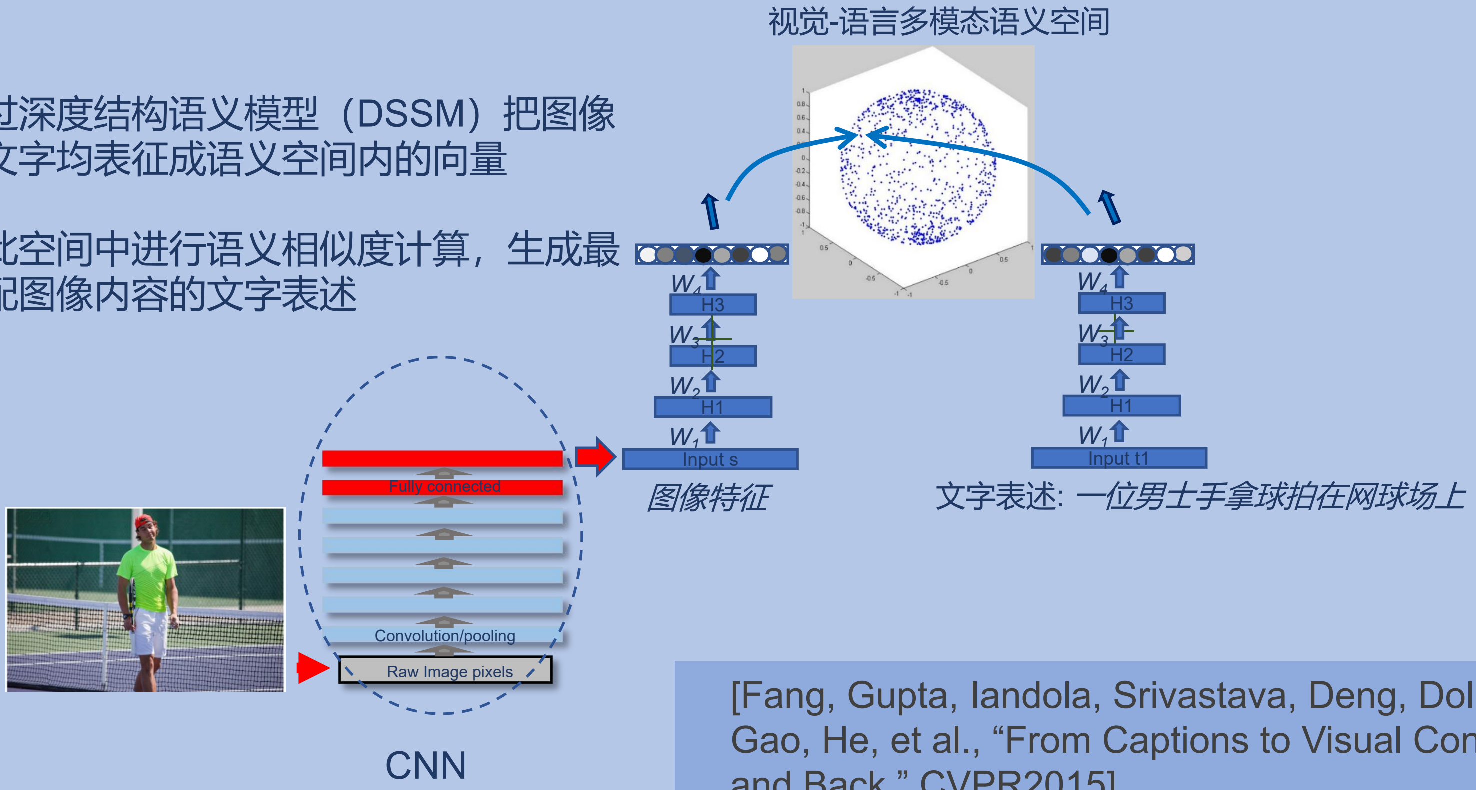


2018

建立多模态语义空间：跨模态表征学习

通过深度结构语义模型（DSSM）把图像和文字均表征成语义空间内的向量

在此空间中进行语义相似度计算，生成最匹配图像内容的文字表述



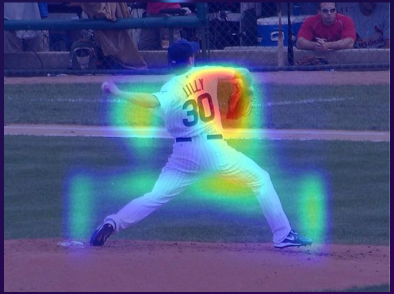
[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, et al., “From Captions to Visual Concepts and Back,” CVPR2015]

2018

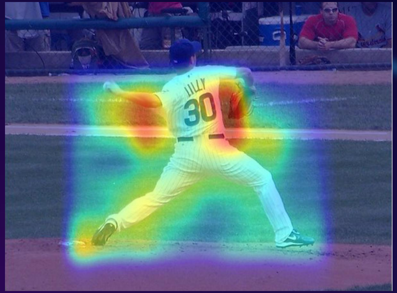
图像描述：理解图像，用语言表达

a baseball player throwing a ball

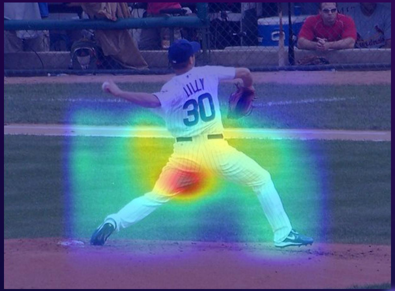
“一个棒球运动员在扔一个球。”



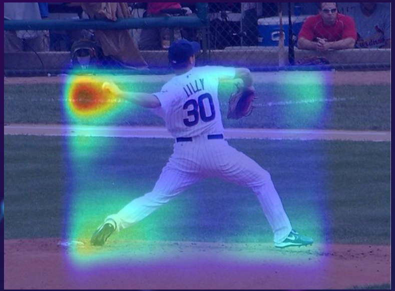
一个棒球



一个棒球运动员



一个棒球运动员在扔

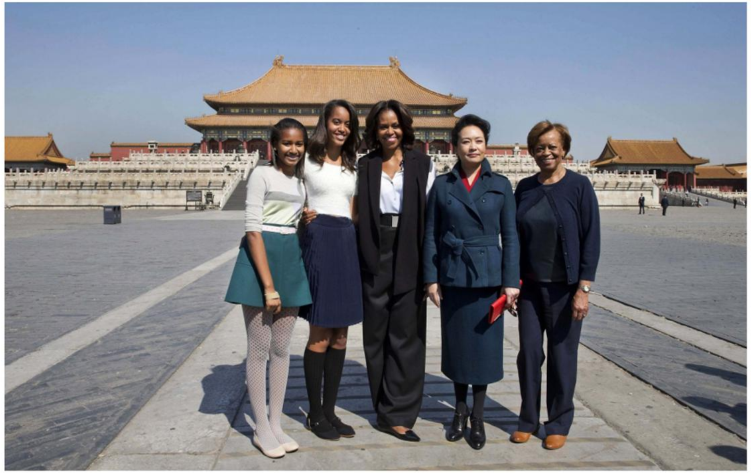


一个棒球运动员在扔一个球

2018

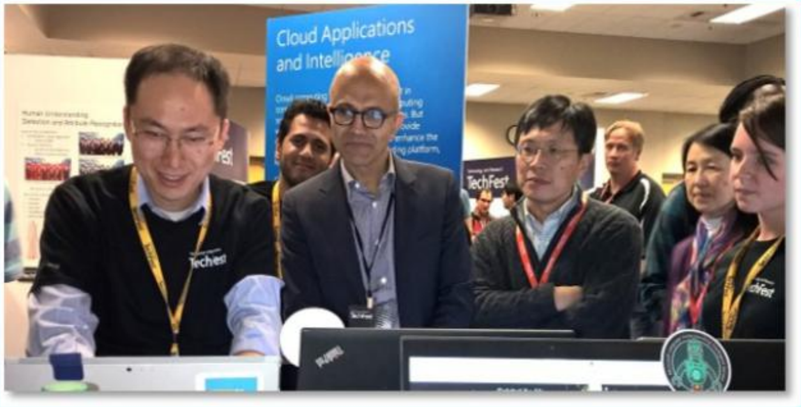
[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He,
et al., “From Captions to Visual Concepts and Back,” CVPR2015]

与实体知识融合



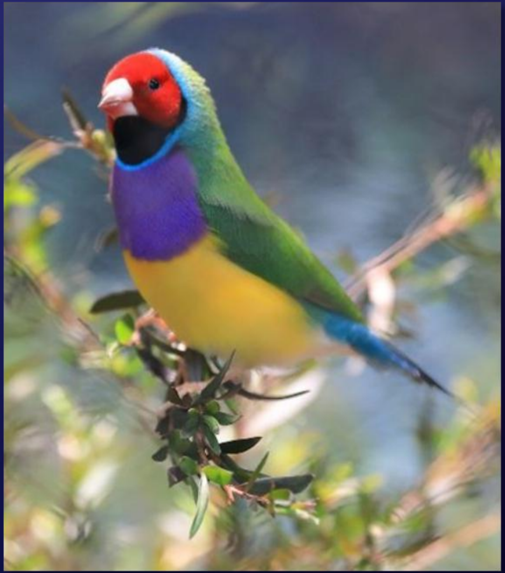
Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background.

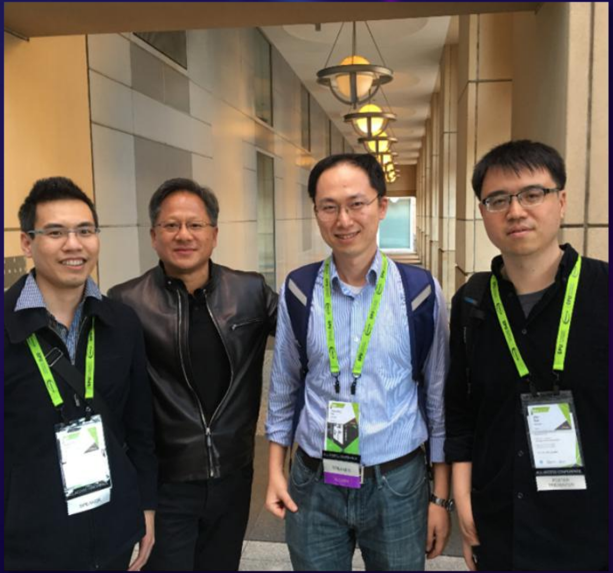
I think it's Satya Nadella, Harry Shum posing for the camera.

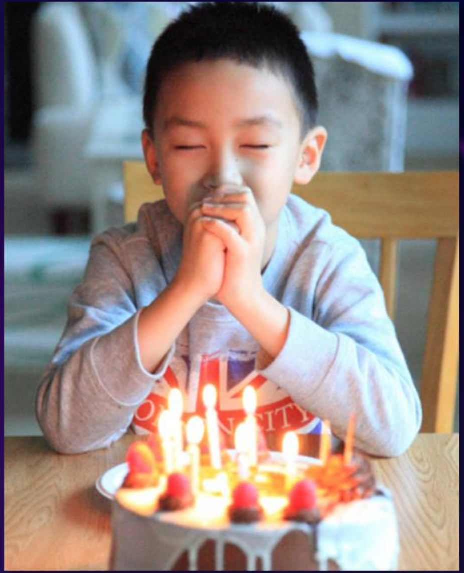


[Guo, Zhang, Hu, He, Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition", ECCV 2016; Tran, He, Zhang, Sun, et al., "Rich image captioning in the wild," CVPR DeepVision Workshop 2016]

图片描述机器人 [http://captionbot.ai]







"A colorful bird perched on a tree branch."

一只多彩的小鸟在树枝上鸣叫。

"Jen-Hsun Huang, Xiaodong He, Jian Sun et al., that are posing for a picture."

黄仁勋,何晓冬,孙剑等合影留念。

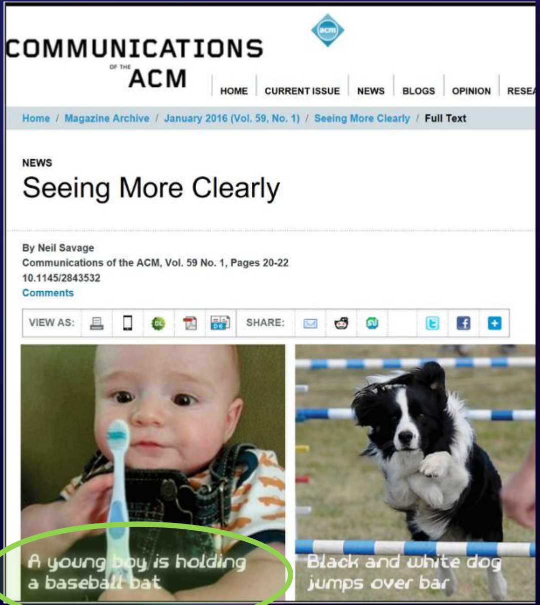
"A little boy sitting in front of a birthday cake and he seems happy." —位小男孩坐在生日蛋糕前，看起来很高兴。

2018

控制语言的语义生成

2018

Example in CACM 2016



COMMUNICATIONS OF THE ACM

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEA

Home / Magazine Archive / January 2016 (Vol. 59, No. 1) / Seeing More Clearly / **Full Text**

NEWS

Seeing More Clearly

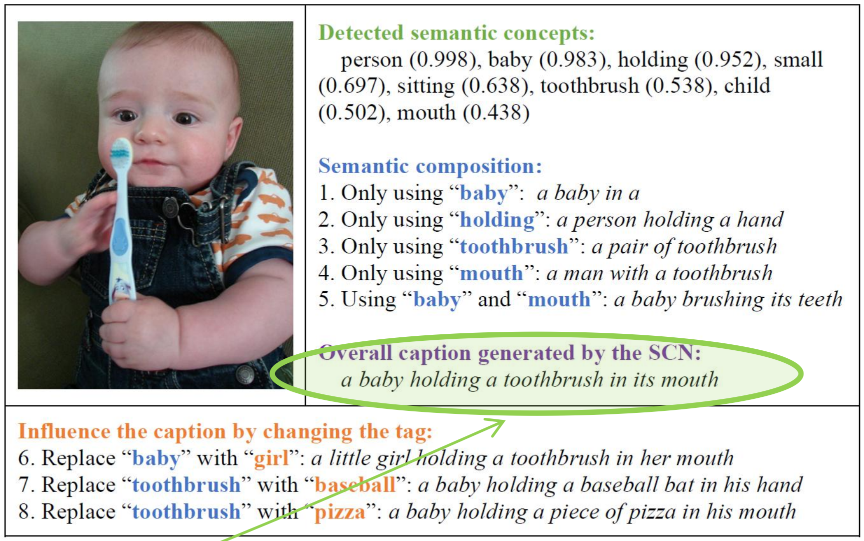
By Neil Savage
Communications of the ACM, Vol. 59 No. 1, Pages 20-22
10.1145/2843532
Comments

VIEW AS: SHARE:

A young boy is holding a baseball bat

Black and white dog jumps over bar

Semantic Compositional Networks



Detected semantic concepts:
person (0.998), baby (0.983), holding (0.952), small (0.697), sitting (0.638), toothbrush (0.538), child (0.502), mouth (0.438)

Semantic composition:
1. Only using "**baby**": *a baby in a*
2. Only using "**holding**": *a person holding a hand*
3. Only using "**toothbrush**": *a pair of toothbrush*
4. Only using "**mouth**": *a man with a toothbrush*
5. Using "**baby**" and "**mouth**": *a baby brushing its teeth*

Overall caption generated by the SCN:
a baby holding a toothbrush in its mouth

Influence the caption by changing the tag:
6. Replace "**baby**" with "**girl**": *a little girl holding a toothbrush in her mouth*
7. Replace "**toothbrush**" with "**baseball**": *a baby holding a baseball bat in his hand*
8. Replace "**toothbrush**" with "**pizza**": *a baby holding a piece of pizza in his mouth*

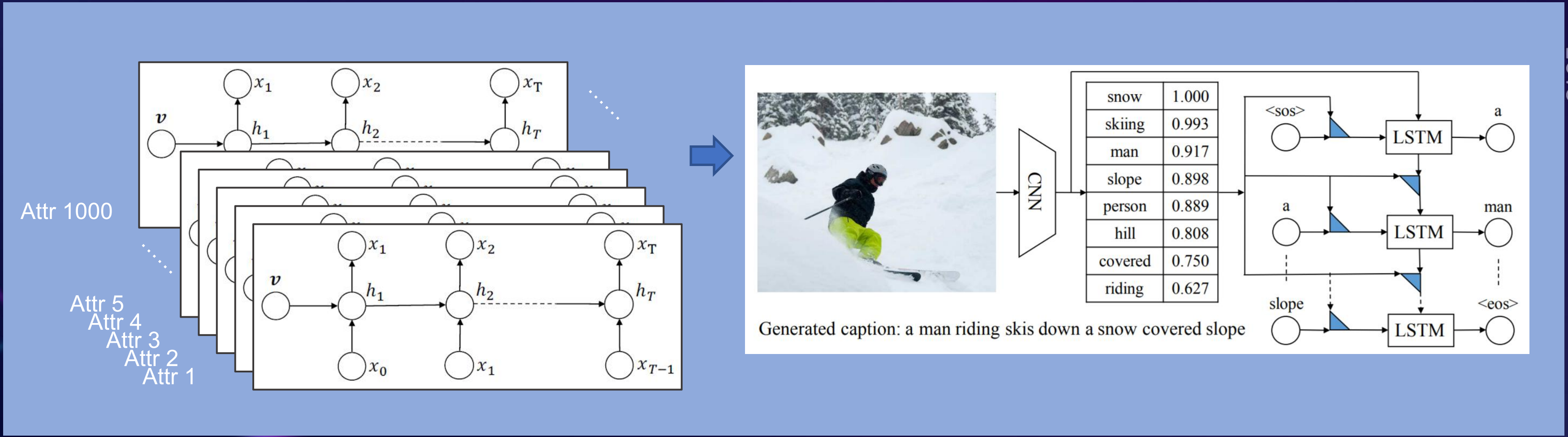
[Gan, Gan, He, Gao, Deng, "Semantic Compositional Networks," CVPR17]

语义组合网络

Semantic Compositional Networks (SCN)

A very wide Model (as wide as 1000 LSTM slices)

- Conceptually, learn 1000 LSTMs, one for each semantic attributes.
- Combine these 1000 LSTMs, weighted by attributes' likelihood.
- Run tensor decomposition to reduce #parameters to fit in GPU



snow	1.000
skiing	0.993
man	0.917
slope	0.898
person	0.889
hill	0.808
covered	0.750
riding	0.627

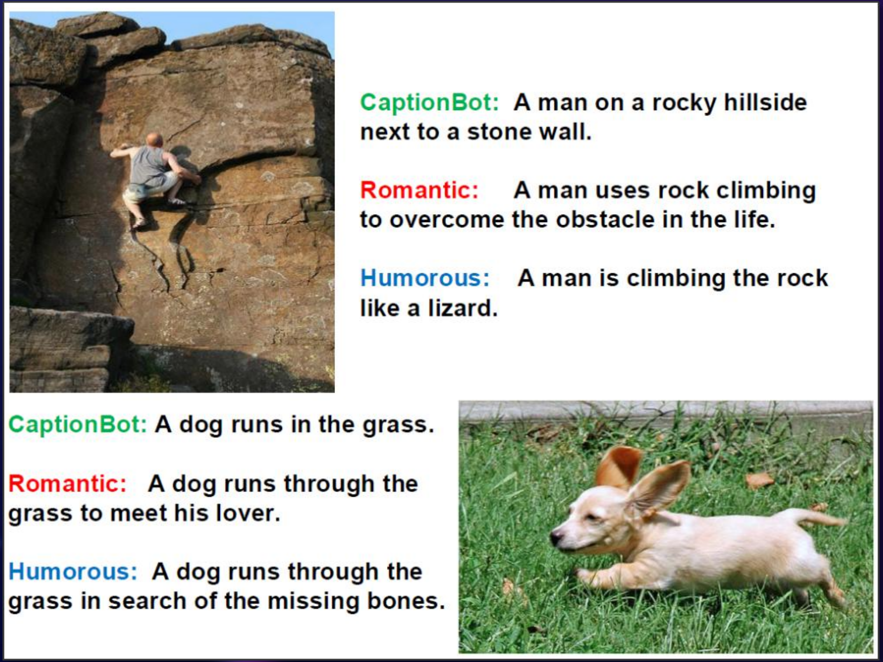
Generated caption: a man riding skis down a snow covered slope

2018

[Gan, Gan, He, Gao, Deng, "*SCN*", CVPR17]

表达情感和风格

让AI用语言来表达浪漫或者幽默的风格 - StyleNet



CaptionBot: A man on a rocky hillside next to a stone wall.

Romantic: A man uses rock climbing to overcome the obstacle in the life.

Humorous: A man is climbing the rock like a lizard.

CaptionBot: A dog runs in the grass.

Romantic: A dog runs through the grass to meet his lover.

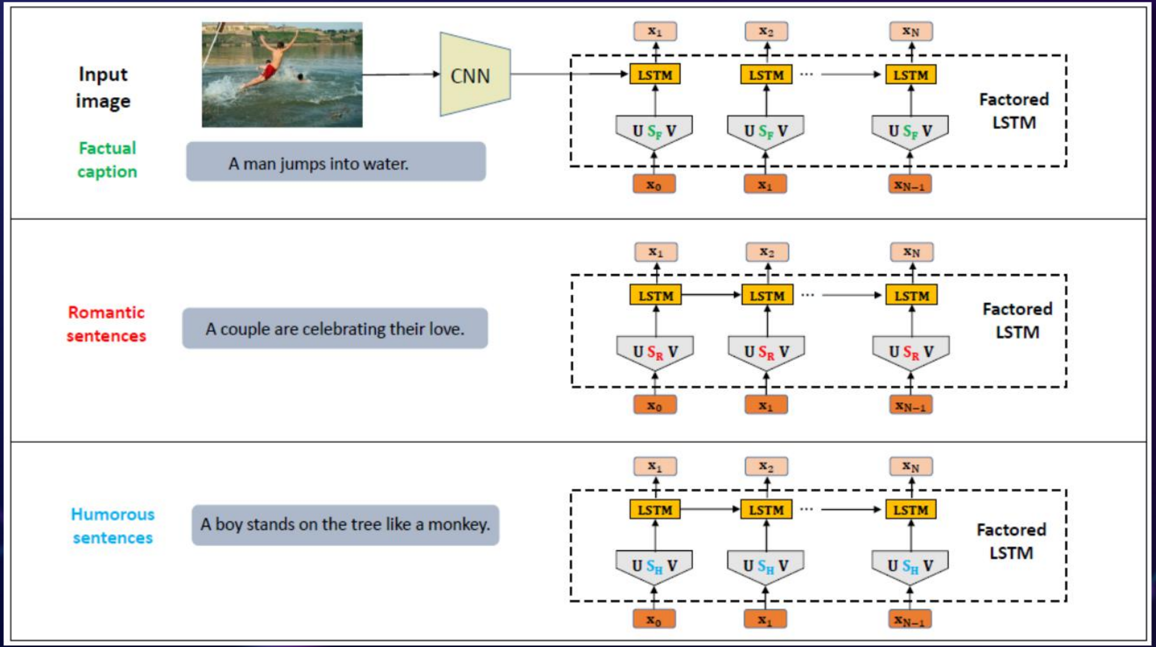
Humorous: A dog runs through the grass in search of the missing bones.

[Gan, Gan, He, Gao, Deng, “*StyleNet*”, CVPR2017]

2018

表达情感和风格

让AI用语言来表达浪漫或者幽默的风格 - StyleNet

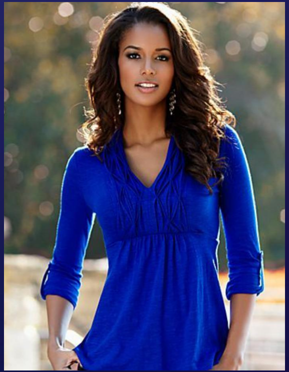


[Gan, Gan, He, Gao, Deng, "*StyleNet*", CVPR2017]

2018

生成带情感的语言

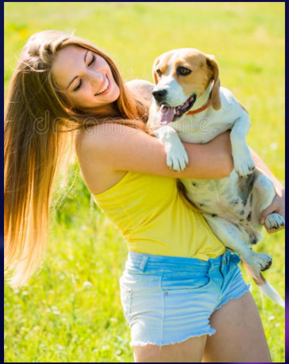
让AI在语言表达中加入情感



分类：　户外，女士

语义：　一位穿着蓝色T恤的女士

情感+：美丽得像一位天使！



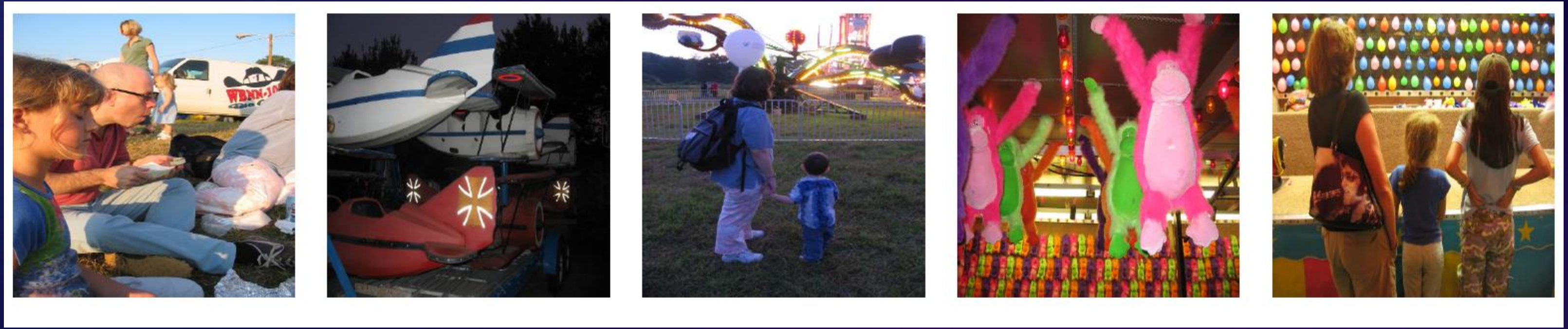
分类：　女士，小狗

语义：　一位女士和一只狗在相机前摆姿势

情感+：啊真可爱，我是说这只小狗耶 ☺

看图生成故事（Storytelling）

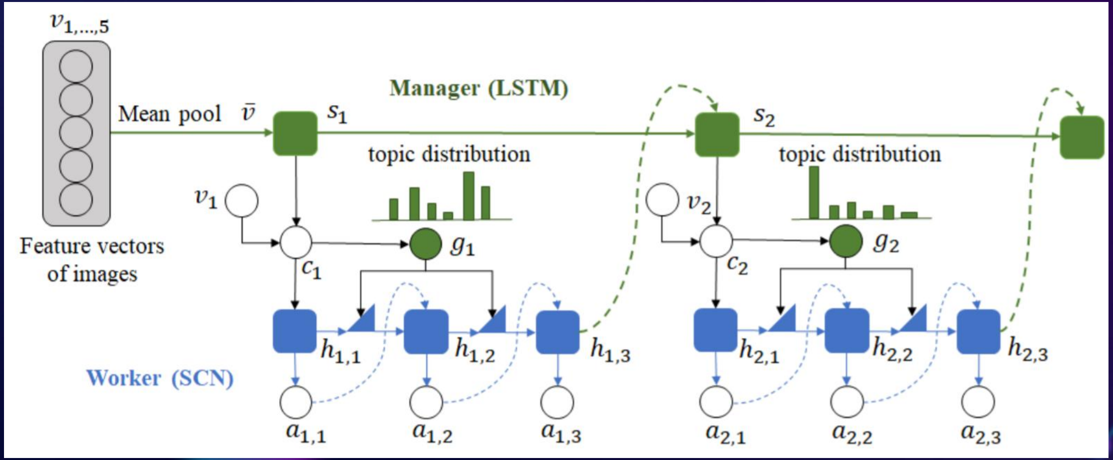
看一组照片写日记



孩子们今天去参加了一个嘉年华会。那有很多不一样的酷酷的东西。有些孩子在玩空中活动。这里还有巨大的恐龙。一天结束了，孩子们玩得很开心。

看图生成故事: 生成多句相关句子

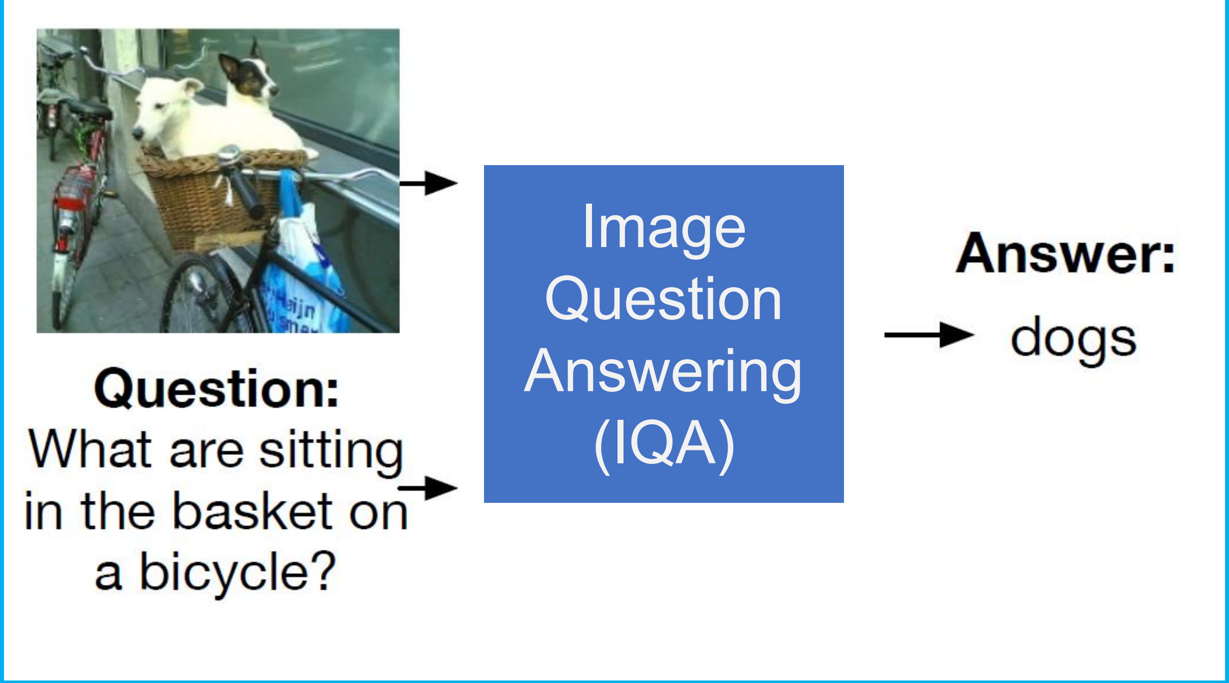
双层生成模型：
上层（Manager）：生成一系列主题（topics）
下层（Worker）　：按主题展开成句子



[Huang, Gan, et al., “Hierarchically structured reinforcement learning for topically coherent visual story generation,” AAAI2019]

视觉-语言多模态机器问答（Visual QA）

- 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)

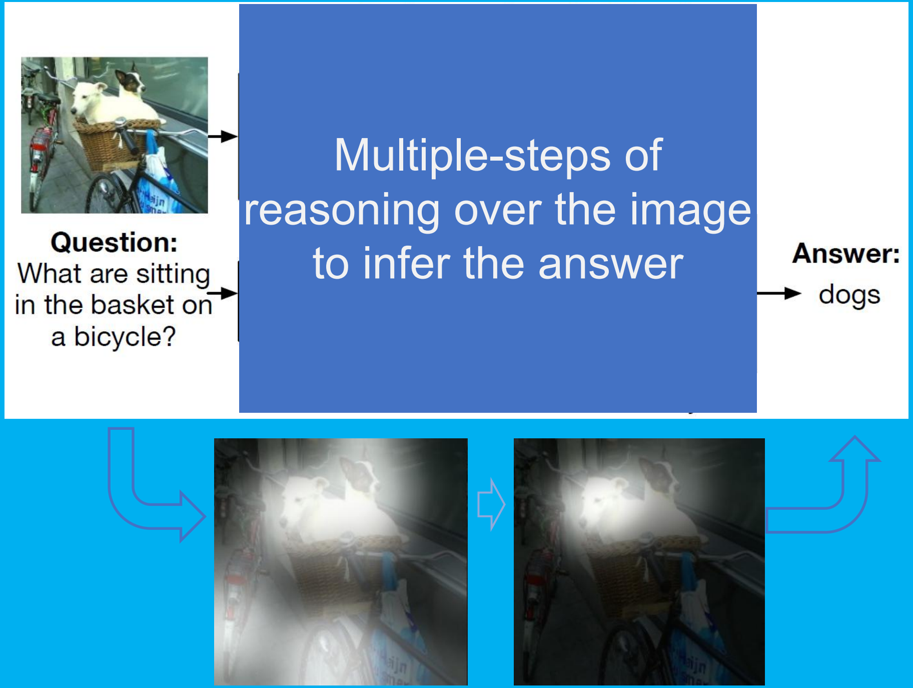
Answer:
dogs

从图片描述到图文问答： 推理能力

To answer a question about a image:

Need to understand subtle relationships among multiple objects

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

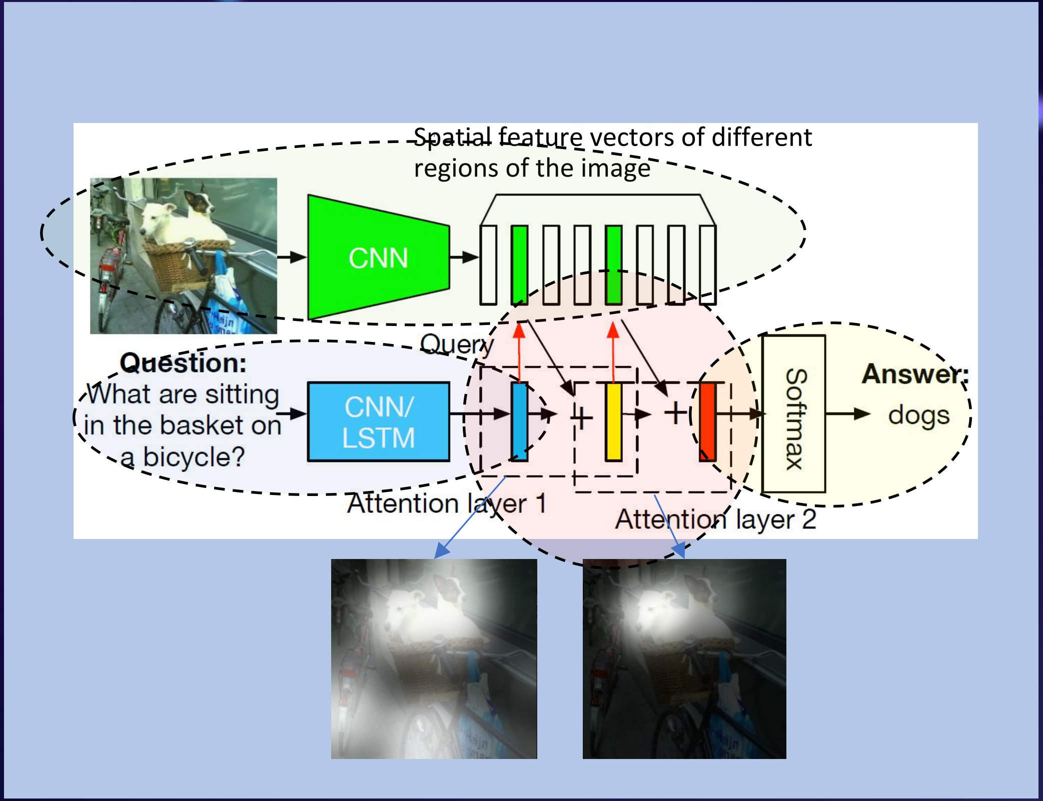


Question:
What are sitting in the basket on a bicycle?
Multiple-steps of reasoning over the image to infer the answer
Answer:
dogs

堆叠注意力网络 (Stacked Attention Net)

SANs perform multi-step reasoning

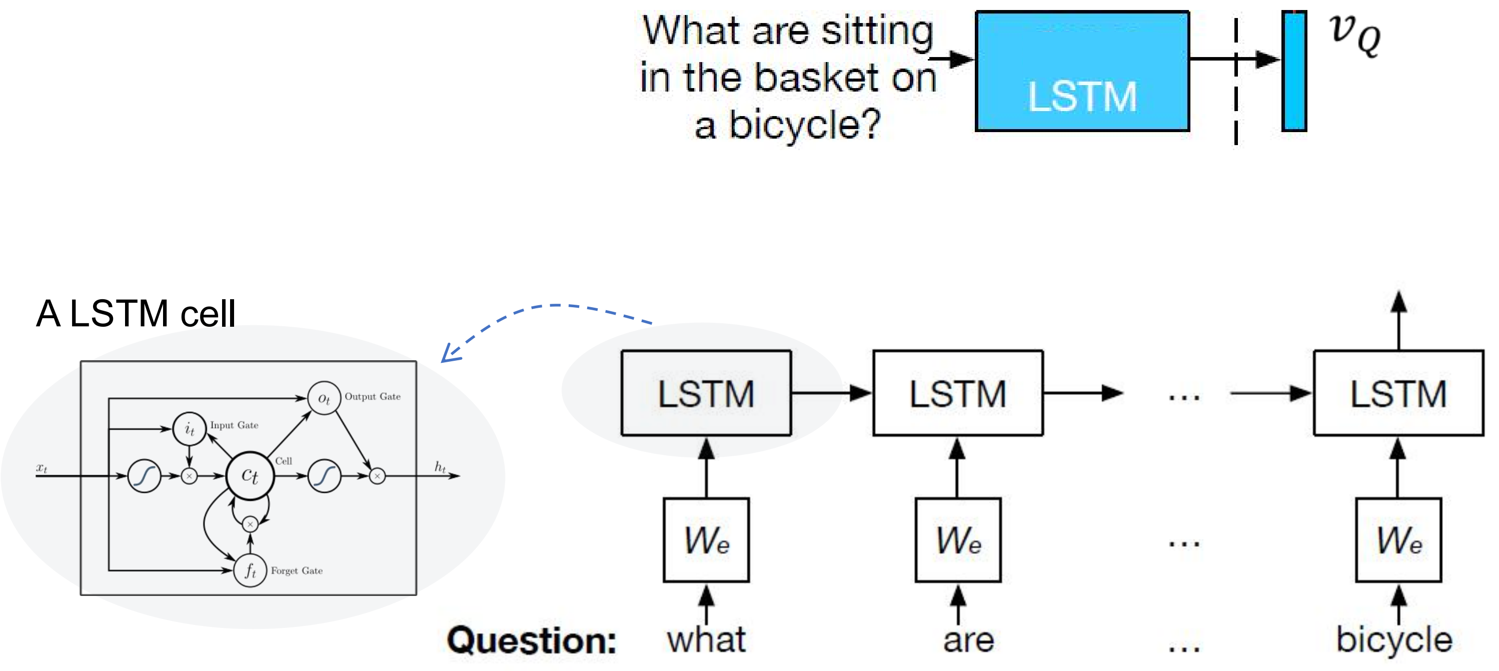
1. Question model
2. Image model
3. Multi-level attention model
4. Answer predictor
5. End-to-end learning using SGD



Spatial feature vectors of different regions of the image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/ LSTM
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

[Yang, He, Gao, Deng, Smola, “Stacked Attention Networks,” CVPR 2016]

提取语言表征


What are sitting in the basket on a bicycle?
LSTM
v_Q
A LSTM cell
x_t
i_t
Input Gate
o_t
Output Gate
c_t
Cell
f_t
Forget Gate
h_t
LSTM
LSTM
...
LSTM
We
We
...
We
Question:
what
are
...
bicycle

提取视觉表征

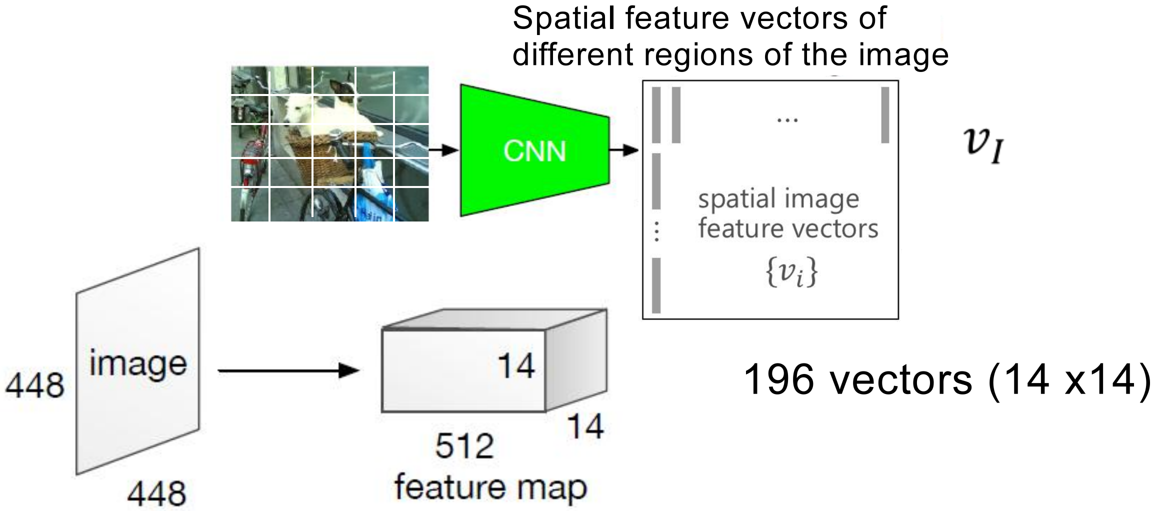
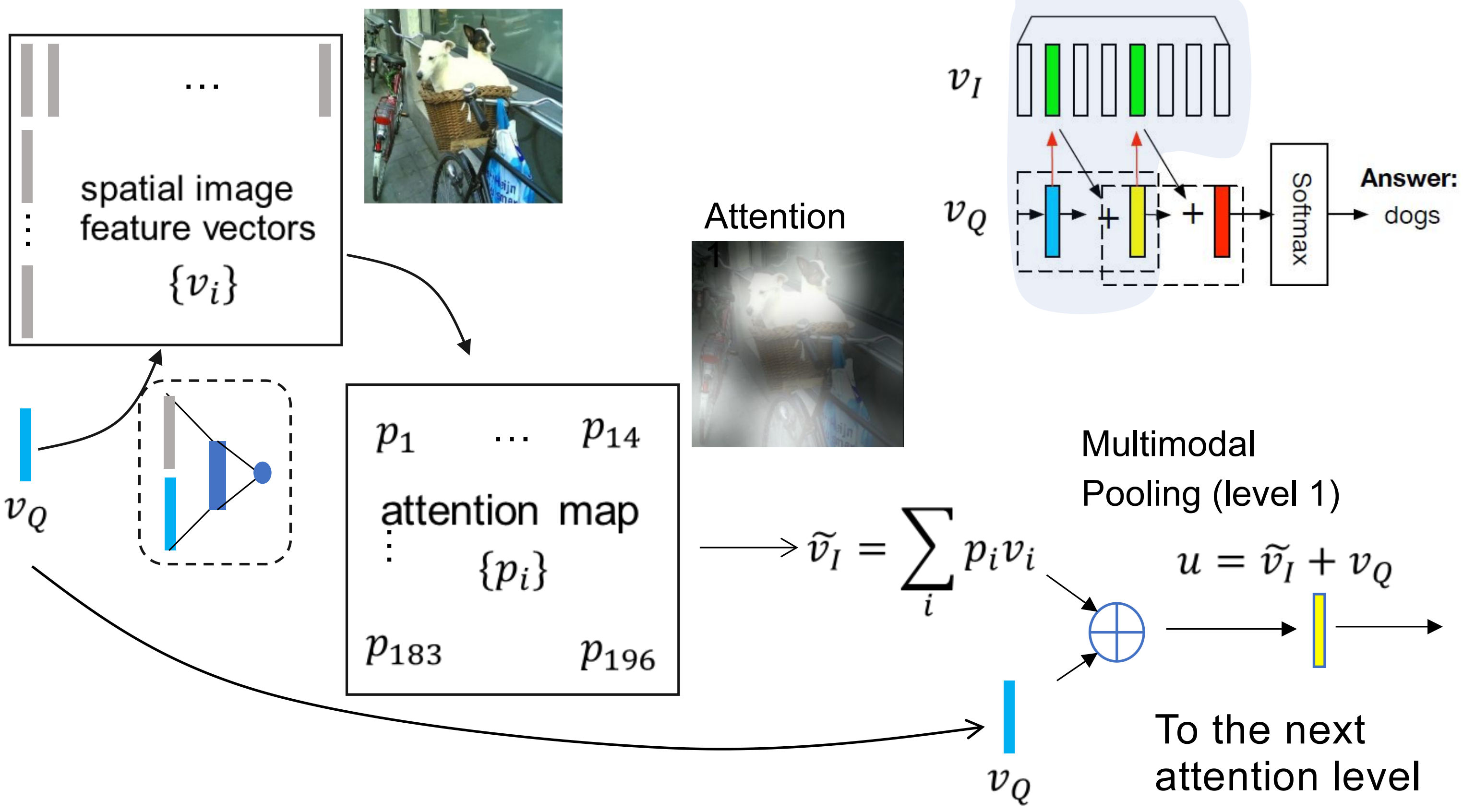

Spatial feature vectors of different regions of the image
CNN
...
spatial image feature vectors
$\{v_i\}$
v_I
196 vectors (14 x14)
448
image
448
14
14
512
feature map

Figure 2: CNN based image model

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

跨模态表征融合与印证(Pooling & Grounding)

第一层注意力

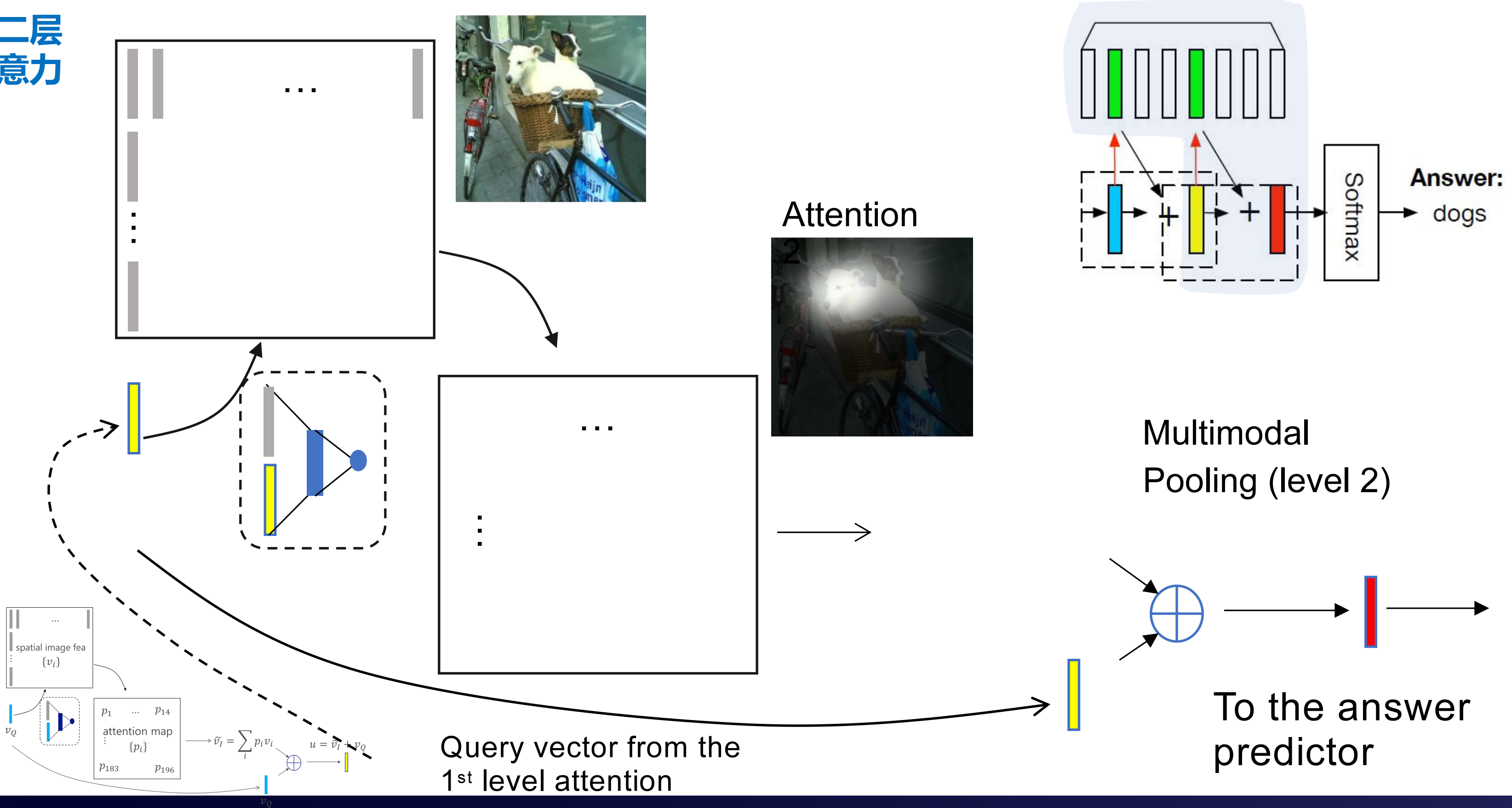


$$\tilde{v}_I = \sum_i p_i v_i$$

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

跨模态表征融合与印证(Pooling & Grounding)

第二层注意力



Attention 2

Multimodal Pooling (level 2)

Query vector from the 1st level attention

To the answer predictor

Bottom-Up and Top-Down Attention（BUTD)

注意力模型的一个新视角

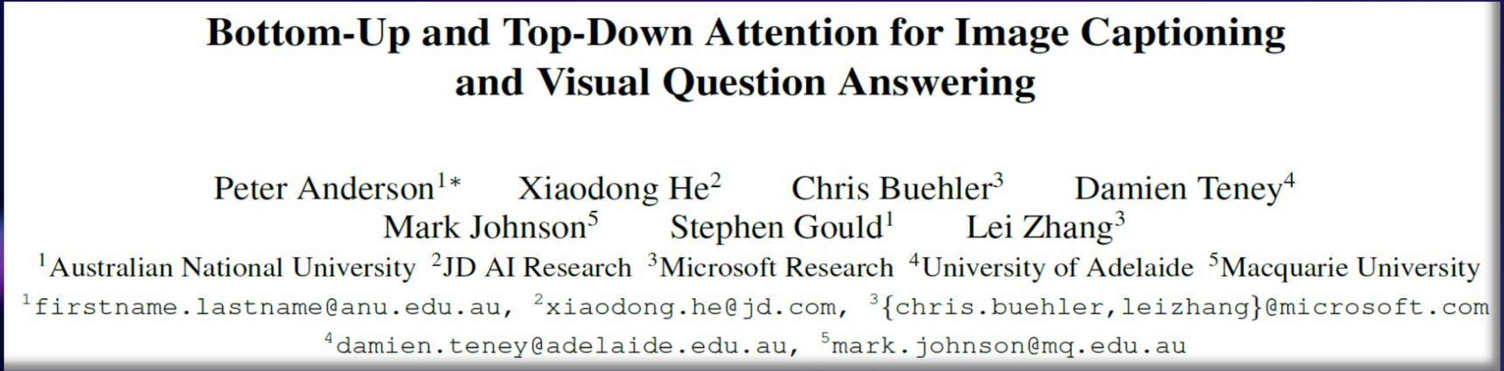
In human visual system, there are two kinds of attentions:

Top-down attention:

proactively initiated by the current task (e.g., look for something)

Bottom-up attention:

spontaneously emerge from visual salient stimuli



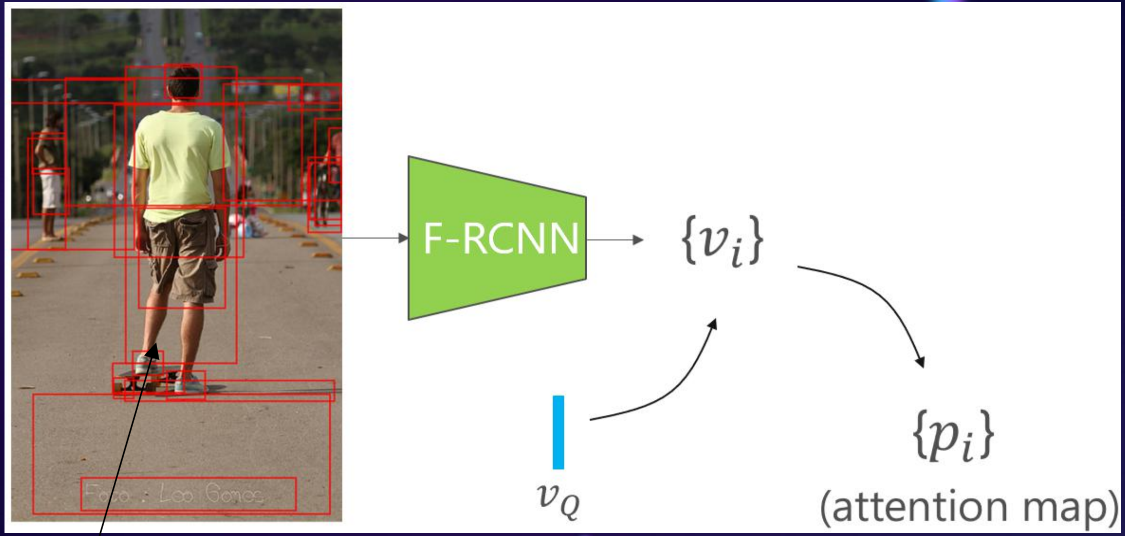
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson[1]* Xiaodong He[2] Chris Buehler[3] Damien Teney[4]
Mark Johnson[5] Stephen Gould[1] Lei Zhang[3]
[1]Australian National University [2]JD AI Research [3]Microsoft Research [4]University of Adelaide [5]Macquarie University
[1]firstname.lastname@anu.edu.au, [2]xiaodong.he@jd.com, [3]{chris.buehler,leizhang}@microsoft.com
[4]damien.teney@adelaide.edu.au, [5]mark.johnson@mq.edu.au

Bottom-Up and Top-Down Attention (BUTD)

Bottom-Up attention:

- Use F-RCNN to detect key objects
- Compute spatial feature vector for each object
- Keep complete visual information for each object



Attend on actual objects, rather than on uniform grid regions like conventional top-down attention

Bottom-Up and Top-Down Attention (BUTD)

Adopt similar terminology to humans' attention system:

- attention mechanisms driven by non visual or task-specific context as 'top-down'
- purely visual feed-forward attention mechanisms as 'bottom-up'.

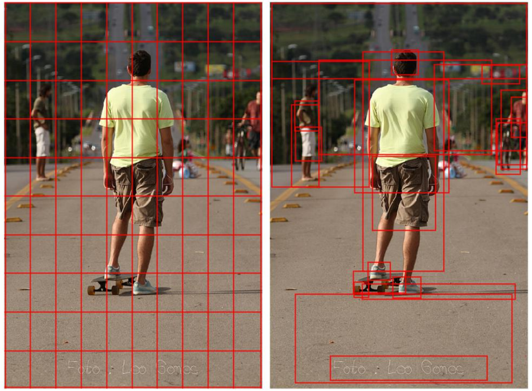
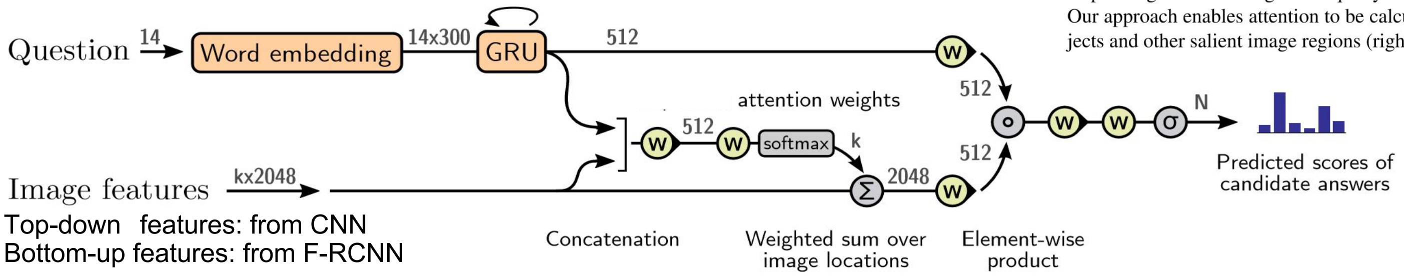


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

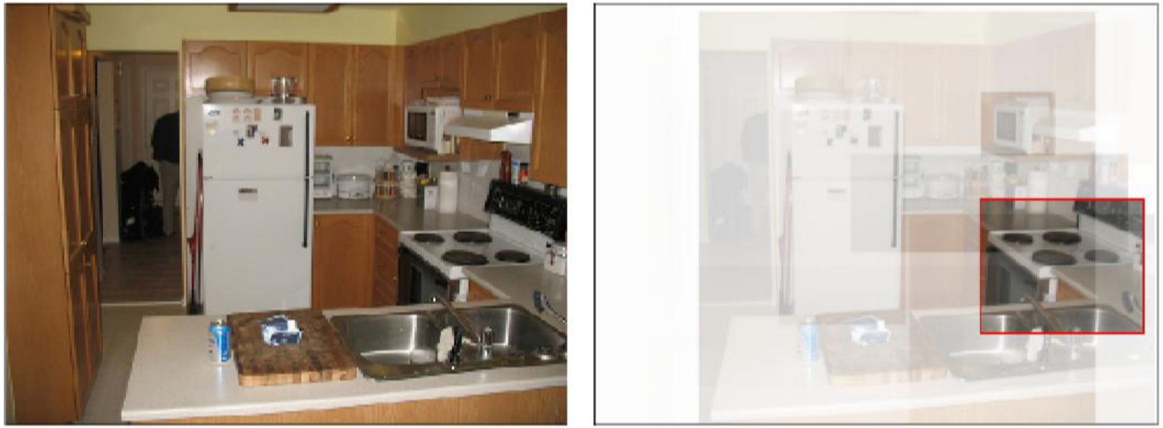
2018

Overall Attention Net for VQA:



Top-down features: from CNN
Bottom-up features: from F-RCNN

Attention Example

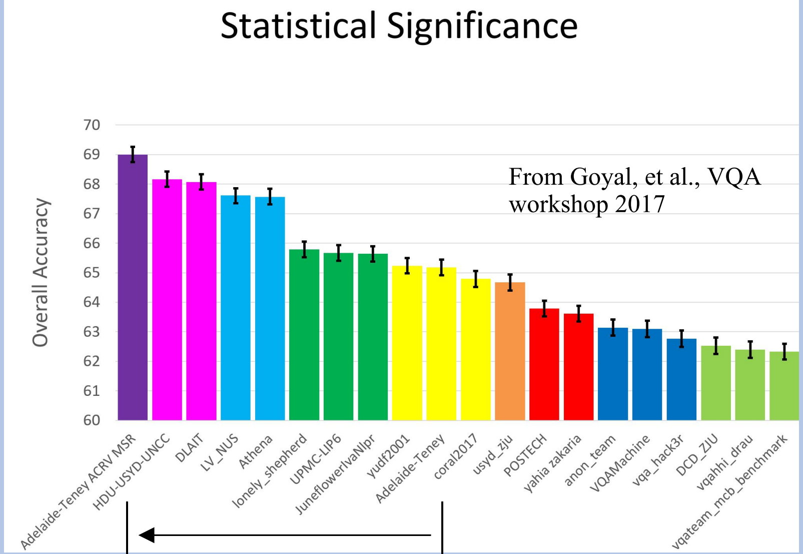


Question: What room are they in? Answer: kitchen

Figure 6. VQA example illustrating attention output. Given the question ‘What room are they in?’, the model focuses on the stovetop, generating the answer ‘kitchen’.

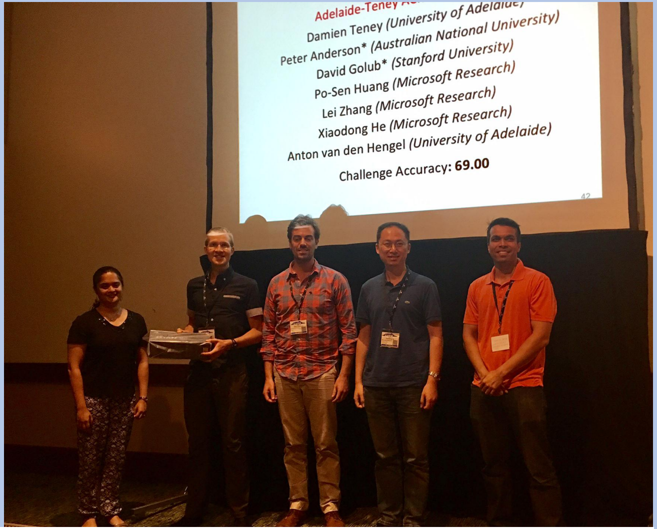
2018

VQA Challenge @ CVPR2017



From Goyal, et al., VQA workshop 2017

Because of **Bottom-up Attention**

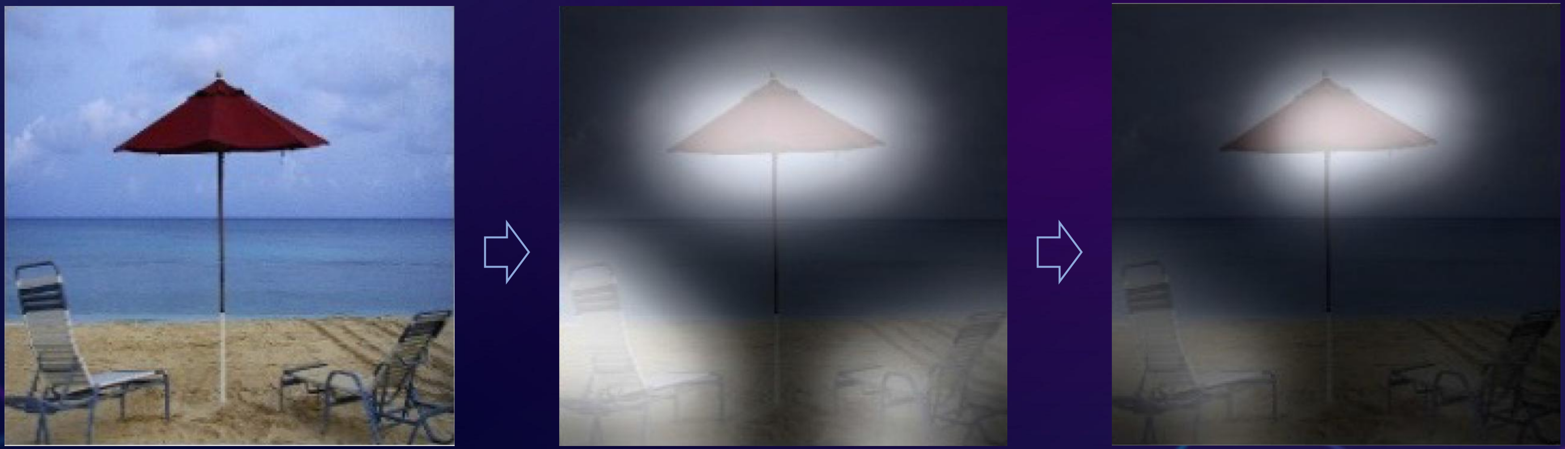

Adelaide-Teney
Damien Teney (University of Adelaide)
Peter Anderson* (Australian National University)
David Golub* (Stanford University)
Po-Sen Huang (Microsoft Research)
Lei Zhang (Microsoft Research)
Xiaodong He (Microsoft Research)
Anton van den Hengel (University of Adelaide)
Challenge Accuracy: 69.00

[1] Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, CVPR18
[2] Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge, CVPR18

此后几乎所有的VQA队伍都使用了“Bottom-Up and Top-Down (BUTD)”注意力模型或其变种。

VQA应答机器人

那两把蓝色椅子之间是什么？



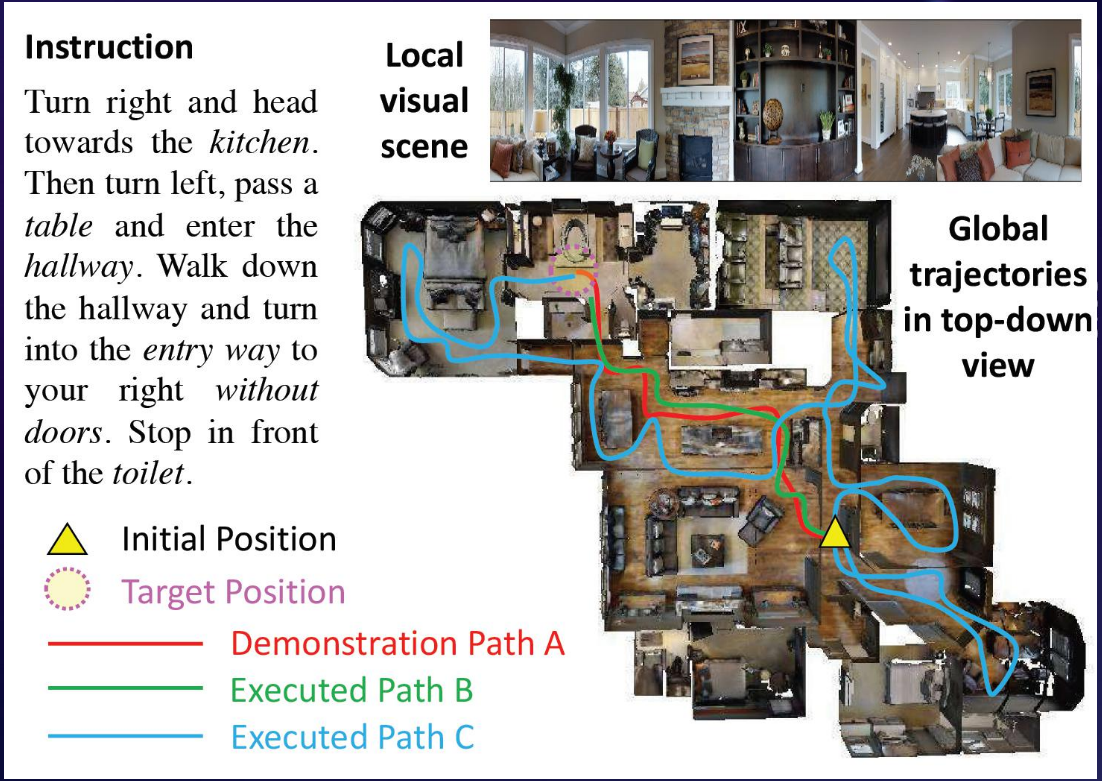
一把伞

【Yang, He, Gao, Deng, Smola, CVPR2016】

2018

视觉-语言多模态导航

结合语言理解和对环境的视觉信息建模，智能代理能按指令从一个地方走到另一个地方



Instruction

Turn right and head towards the *kitchen*. Then turn left, pass a *table* and enter the *hallway*. Walk down the hallway and turn into the *entry way* to your right *without doors*. Stop in front of the *toilet*.

[Anderson et al., CVPR2018;
Wang et al., CVPR 2019]

理解语言，用绘画来表达 (Text-to-Image)

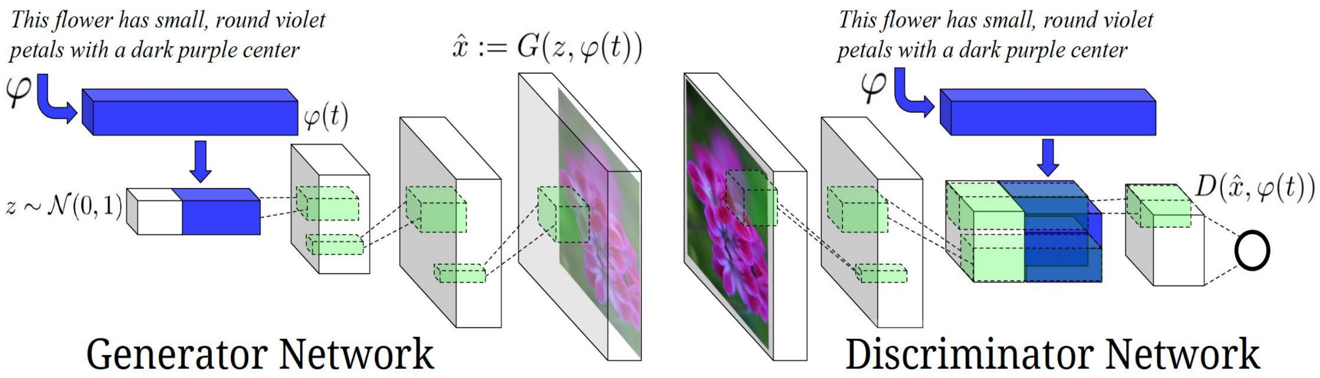
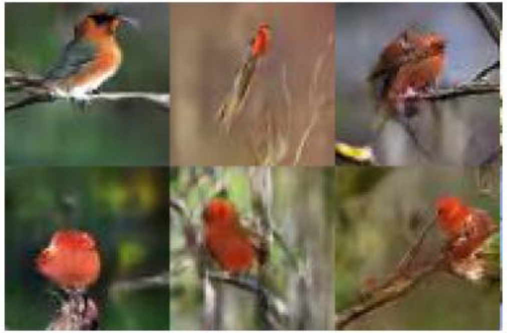

This flower has small, round violet petals with a dark purple center
φ
$\varphi(t)$
$z \sim \mathcal{N}(0,1)$
$\hat{x} := G(z, \varphi(t))$
This flower has small, round violet petals with a dark purple center
φ
$D(\hat{x}, \varphi(t))$
Generator Network
Discriminator Network

Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

this small bird has a pink breast and crown, and black primaries and secondaries.



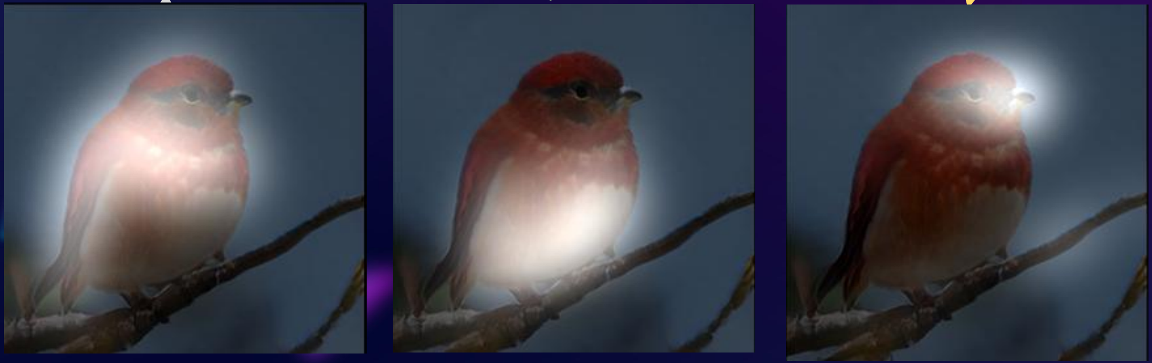
Objective function:

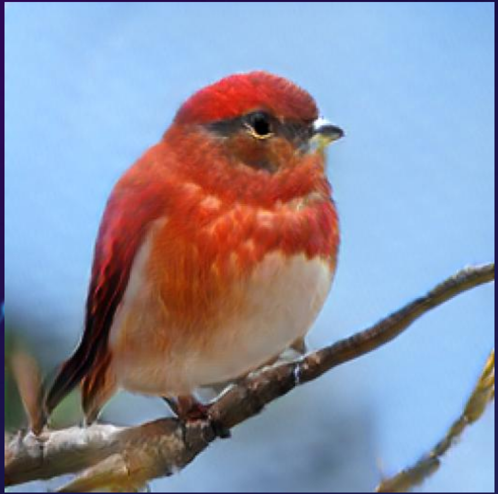
$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

[Reed et al., “Generative adversarial text-to-image synthesis”, ICML2016]

绘画机器人(AttnGAN): 精准理解,精确绘制

一只红羽毛白肚子的短咀小鸟

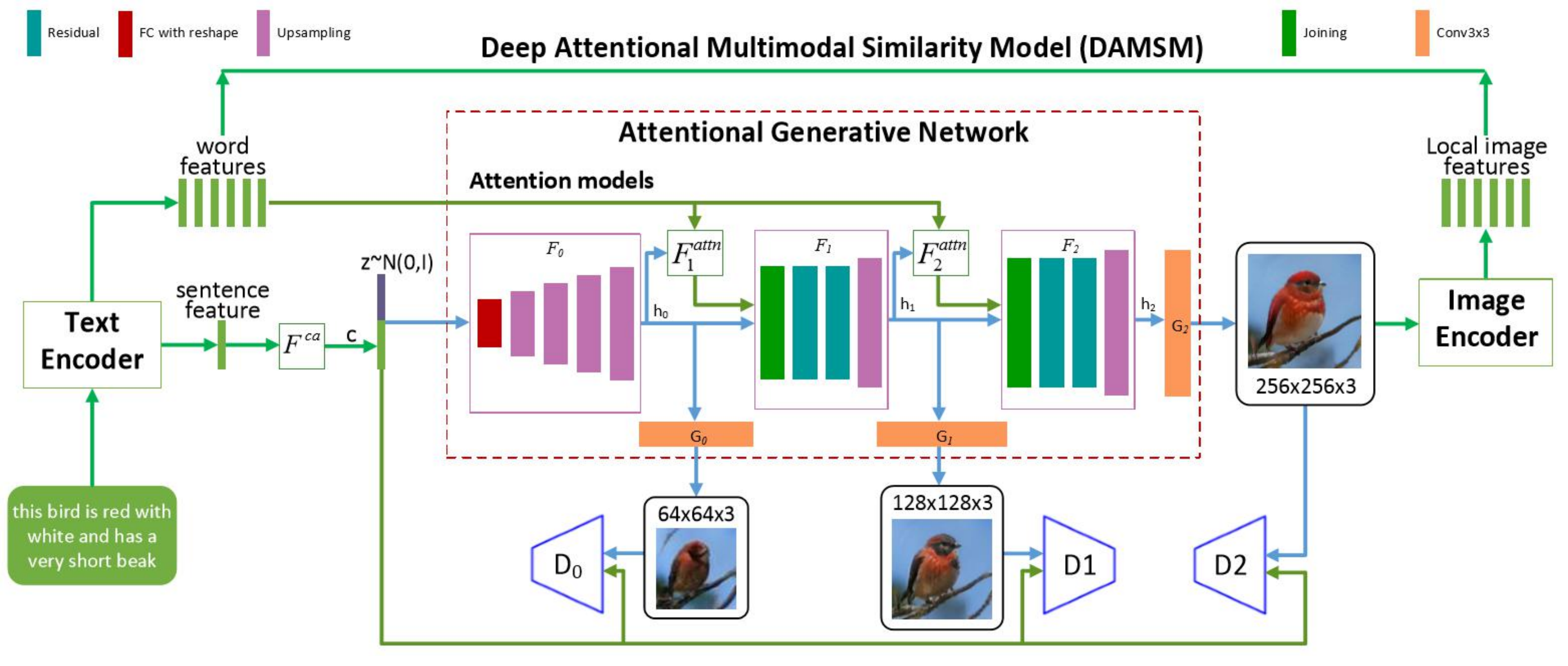




2018

【Xu, Zhang, Huang, Zhang, Gan, Huang, He, “*AttnGAN*,” CVPR2018】

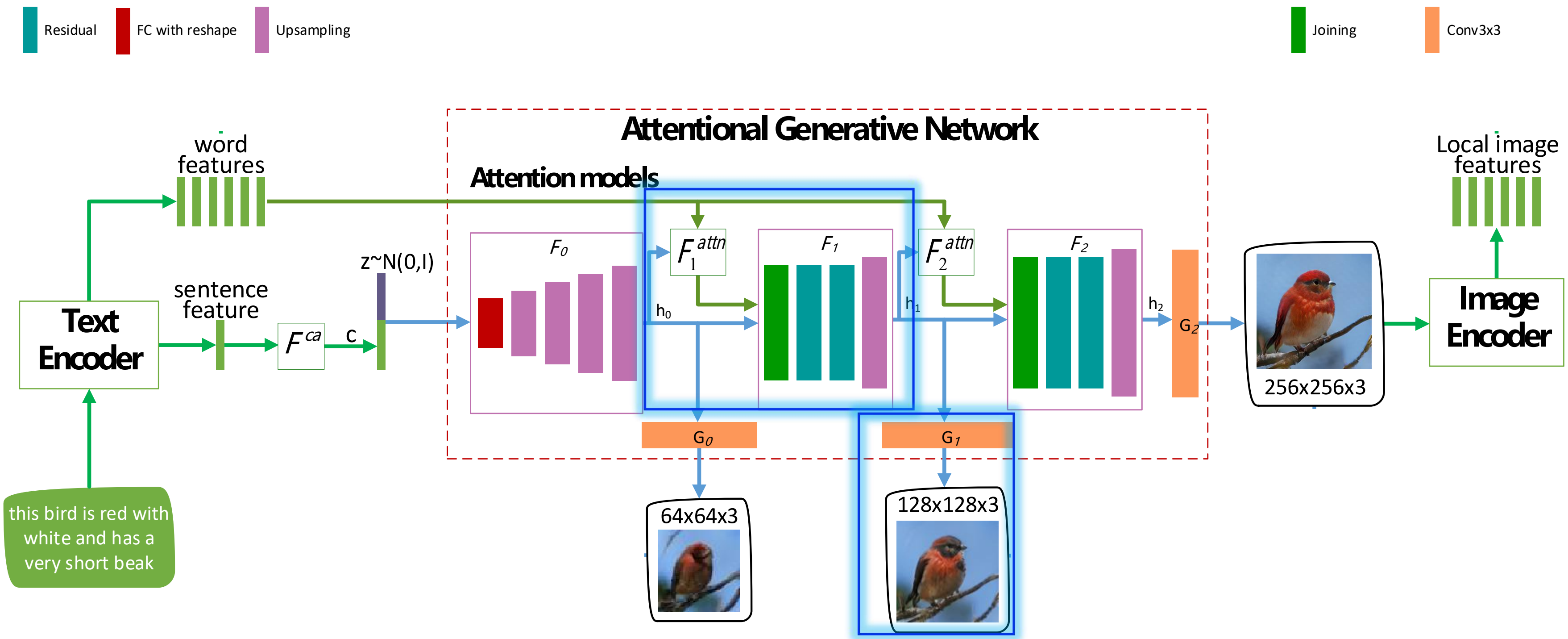
AttnGAN: GAN with Attention



The final objective function: $L = L_{GAN} + \lambda L_{DAMSM}$

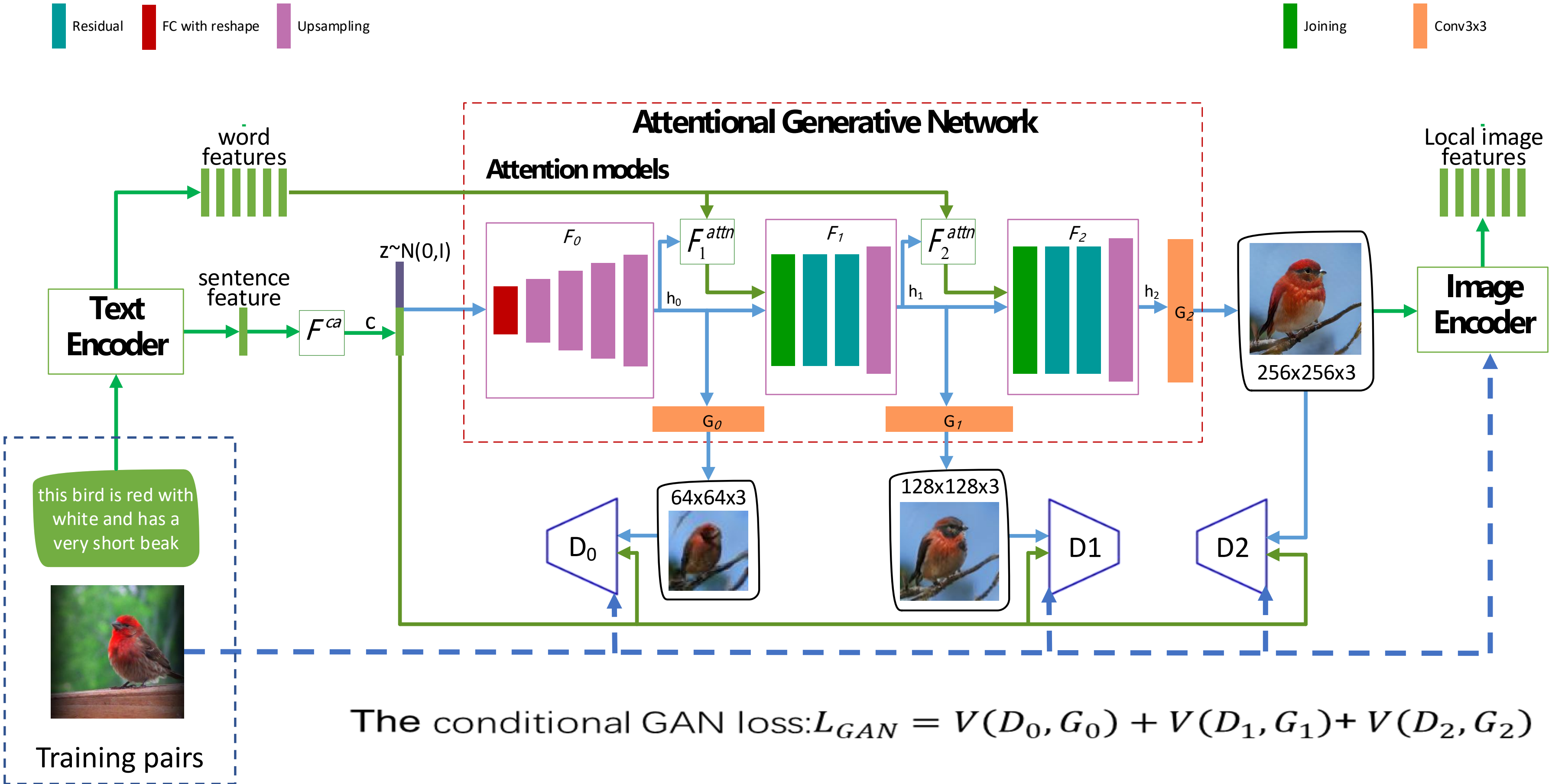
【Xu, Zhang, Huang, Zhang, Gan, Huang, He, “*AttnGAN*,” CVPR2018】

AttnGAN: Attentive Generative Adversarial Networks



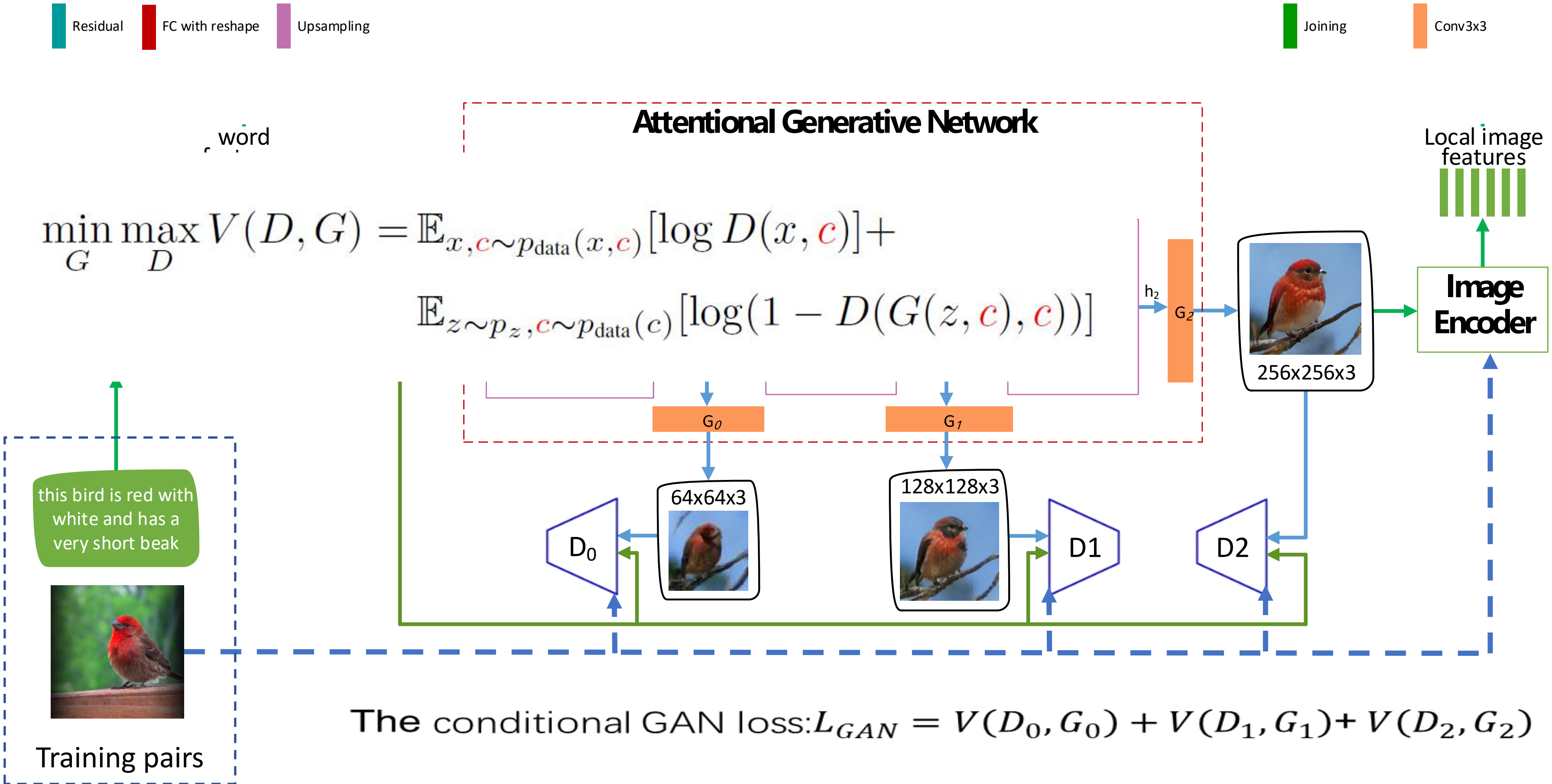
- In following stages, attention models are built.
  - For each region feature of previous generated image, compute its word-context vector.
  - Concatenate previous image region features (e.g., h_0) and word-context vectors to generate the new image.

AttnGAN: Attentive Generative Adversarial Networks



The conditional GAN loss: $L_{GAN} = V(D_0, G_0) + V(D_1, G_1) + V(D_2, G_2)$

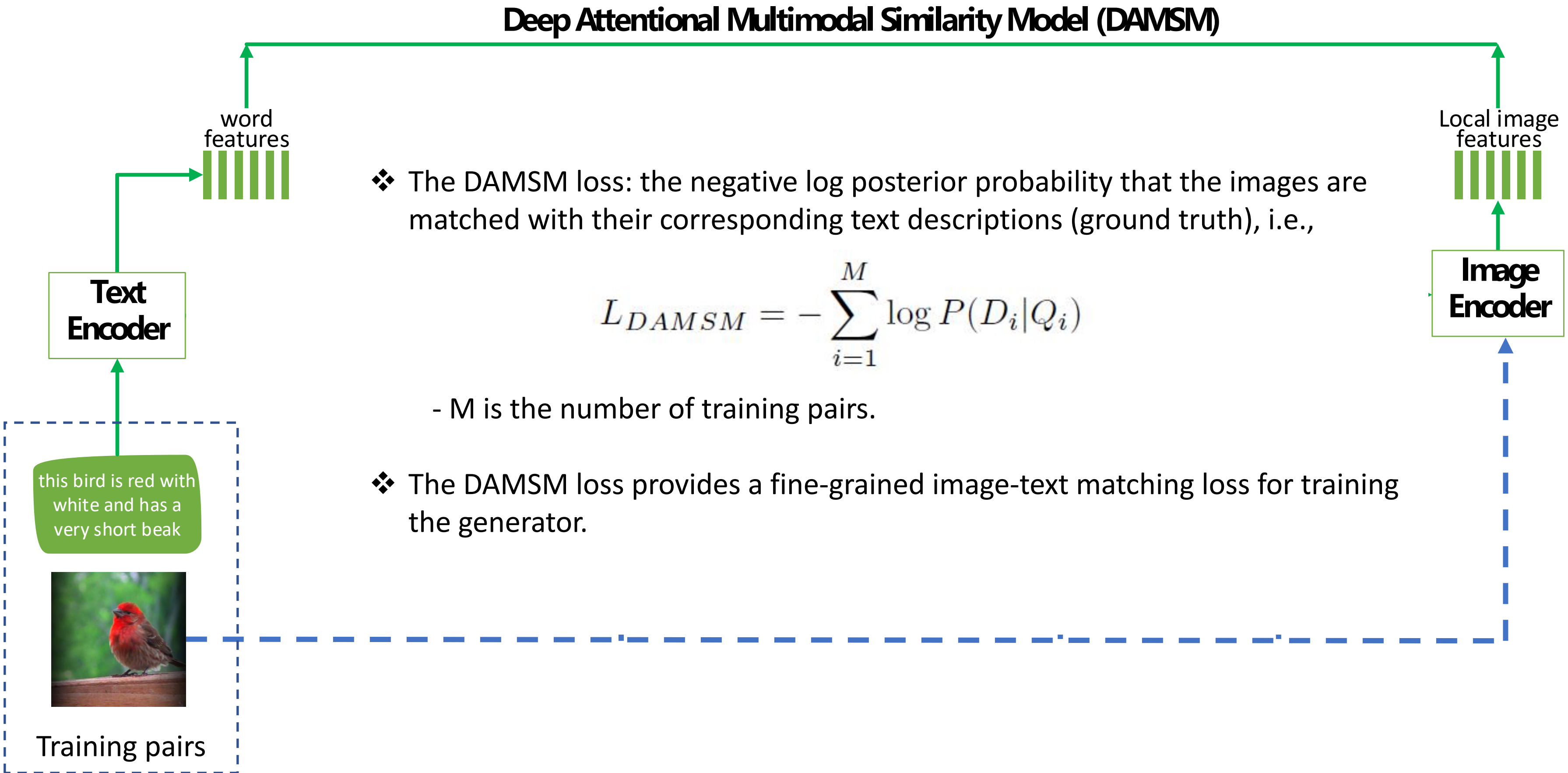
AttnGAN: Attentive Generative Adversarial Networks



$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x,c\sim p_{\text{data}}(x,c)}[\log D(x,c)] + \mathbb{E}_{z\sim p_z, c\sim p_{\text{data}}(c)}[\log(1-D(G(z,c),c))]$$

The conditional GAN loss: $L_{GAN} = V(D_0, G_0) + V(D_1, G_1) + V(D_2, G_2)$

AttnGAN: Attentive Generative Adversarial Networks


Deep Attentional Multimodal Similarity Model (DAMSM)
word features
Local image features
Text Encoder
Image Encoder
this bird is red with white and has a very short beak
Training pairs

❖ The DAMSM loss: the negative log posterior probability that the images are matched with their corresponding text descriptions (ground truth), i.e.,

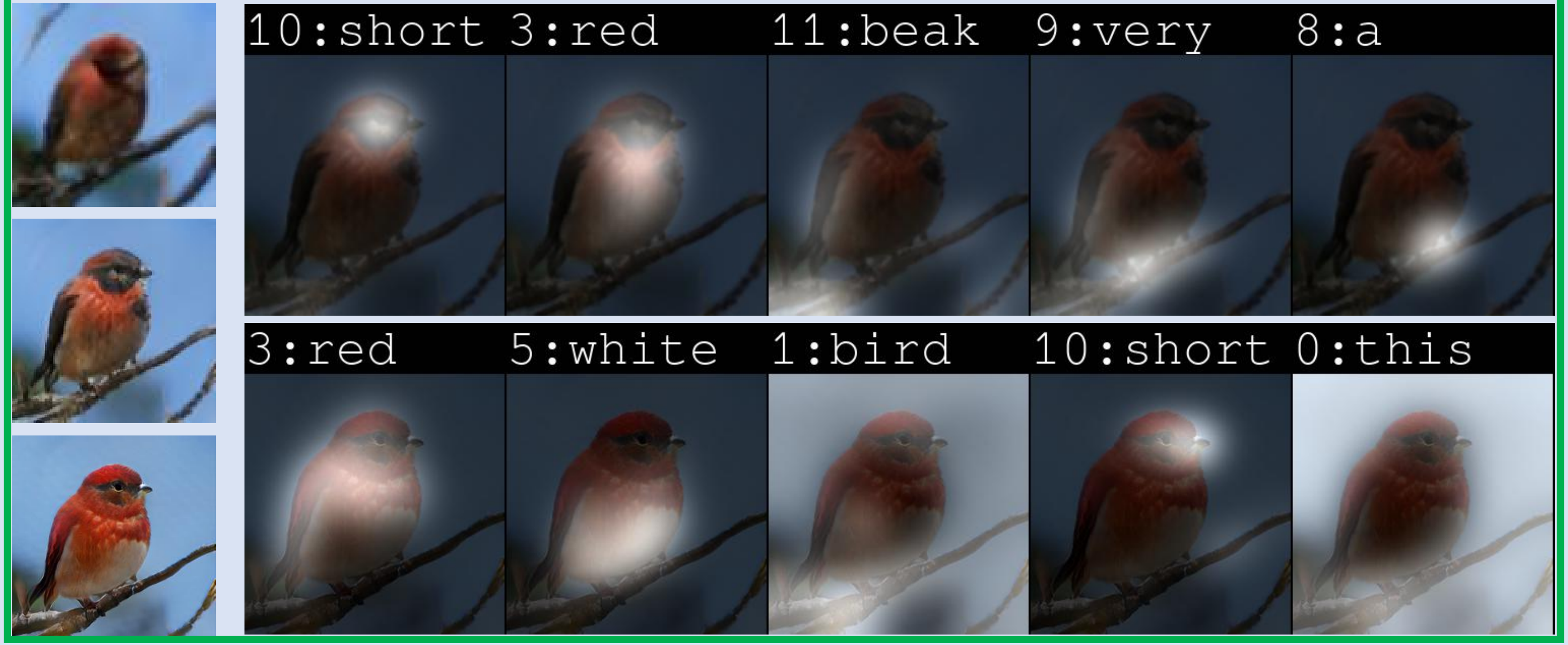
$$L_{DAMSM} = -\sum_{i=1}^{M} \log P(D_i|Q_i)$$

- M is the number of training pairs.

❖ The DAMSM loss provides a fine-grained image-text matching loss for training the generator.

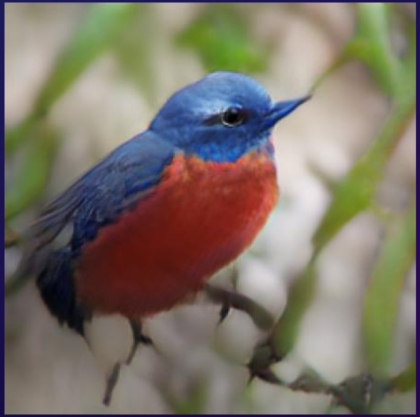
语言-视觉印证：Grounding via Attention

this bird is red with white and has a very short beak.

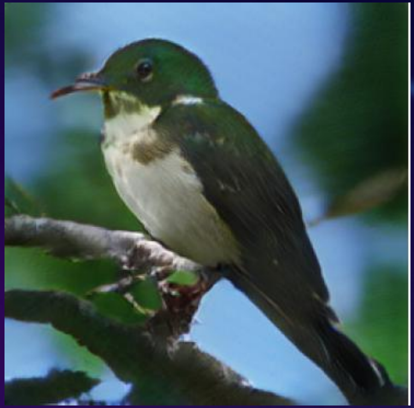


AI的想象: Artificial Imagination

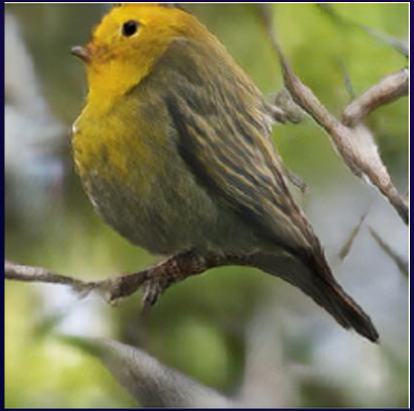
this bird has wings that are blue and has a red belly



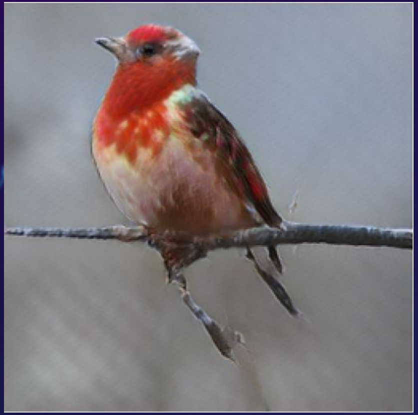
this bird has a green crown black primaries and a white belly



this bird has a yellow crown and a black eye ring that is round



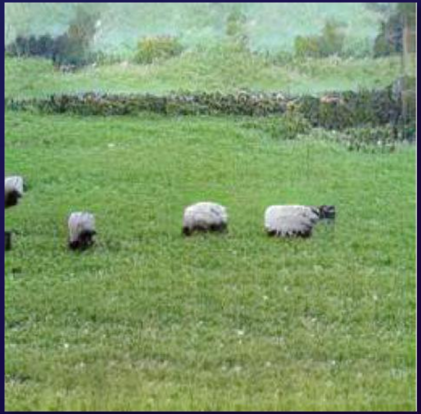
a small red and white bird with a small curved beak



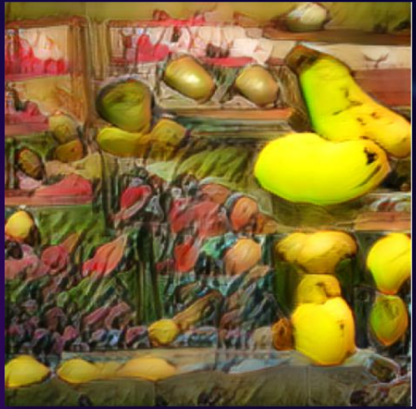
2018

更多绘画

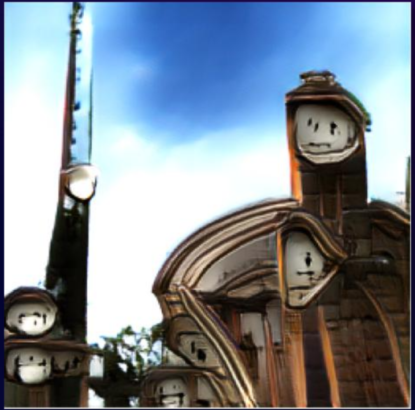
a herd of sheep grazing on a lush green field



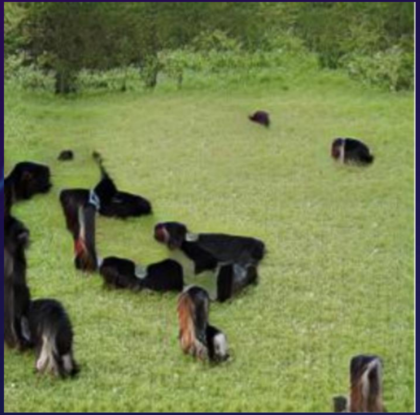
a fruit stand display with bananas and kiwi



an old clock next to a light post in front of a steeple



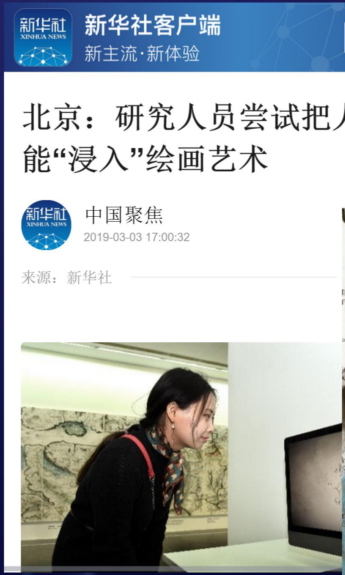
a wild pack of family dogs came running through the yard one day



2018

AI+Art

JDAI与中央美术学院合作



北京：研究人员尝试把人工智能“浸入”绘画艺术

中国聚焦
2019-03-03 17:00:32

来源：新华社

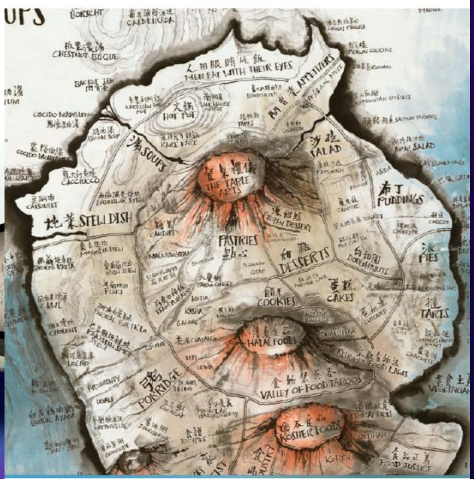
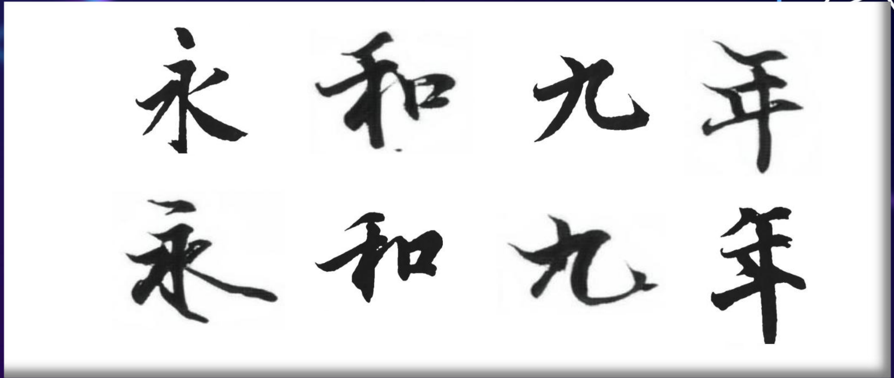
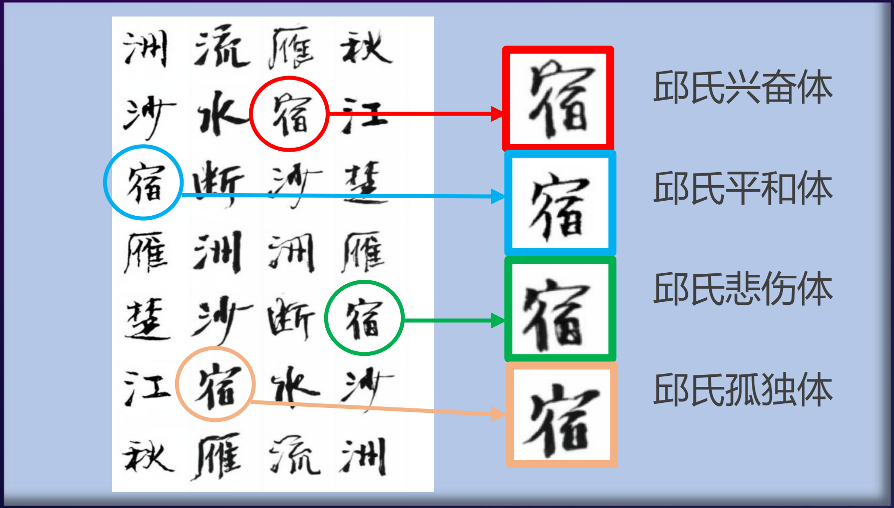


Figure 1: An example of Mind Map

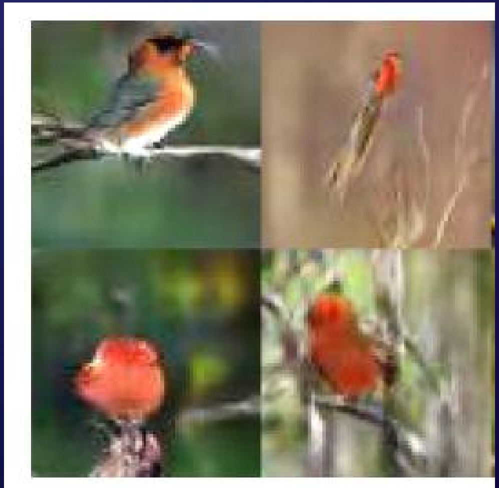
[Liu, et al., 2019]



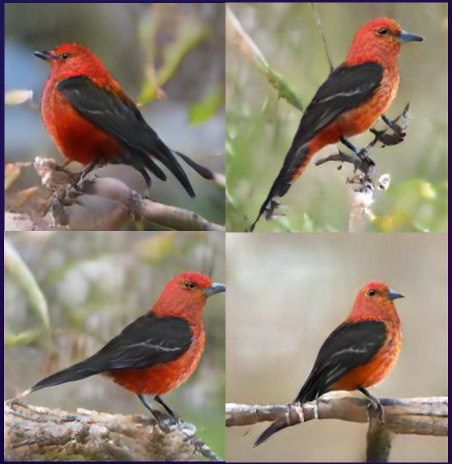


2018

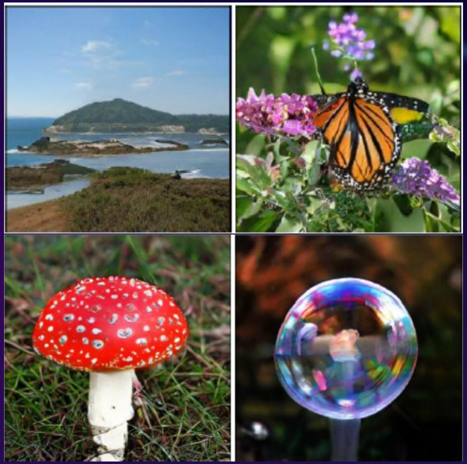
领域进展



[Reed et al., Generative adversarial text-to-image synthesis, ICML, **2016**]



[Xu et al., AttnGAN: Fine-grained text to image generation with Attentional GANs, CVPR **2018**]



[Brock et al., BigGAN: A New State of the Art in Image Synthesis, ICLR **2019**]

2018

多模态智能的下一步

- 融合多个子领域（语言、视觉、语音、知识工程...）
- 跨模态预训练、概念/实体grounding、神经-符号处理机
- 通过复杂跨模态问题驱动基础研究
- 打造更成熟和多元化的实际多模态智能应用

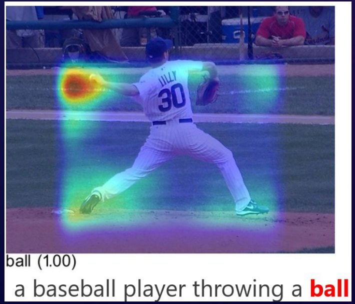
2018

跨越语言和视觉的理解和表达

Universal Chatbot, Digital Assistant, Mixed Reality, Robot w/ Cognition …

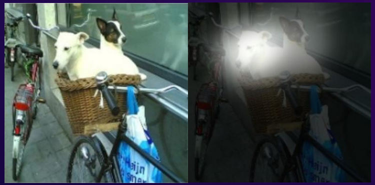
Multimodal Intelligence: perception, reasoning, pooling, grounding and creation across language & vision

Image-to-language



a baseball player throwing a **ball**

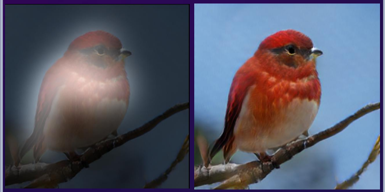
Visual QA/Dialog



Q: what are sitting in the basket on a bicycle?
A: dogs.

Language-to-image

This bird is red with white and has a very short beak



Deep **Representation** Learning / Deep **Attention** Mechanisms

2018

THANKS

xiaodong.he@jd.com

智汇京东 · 开放共赢

JD AI

2018