自然语言理解及多模态智能

何晓冬
IEEE Fellow

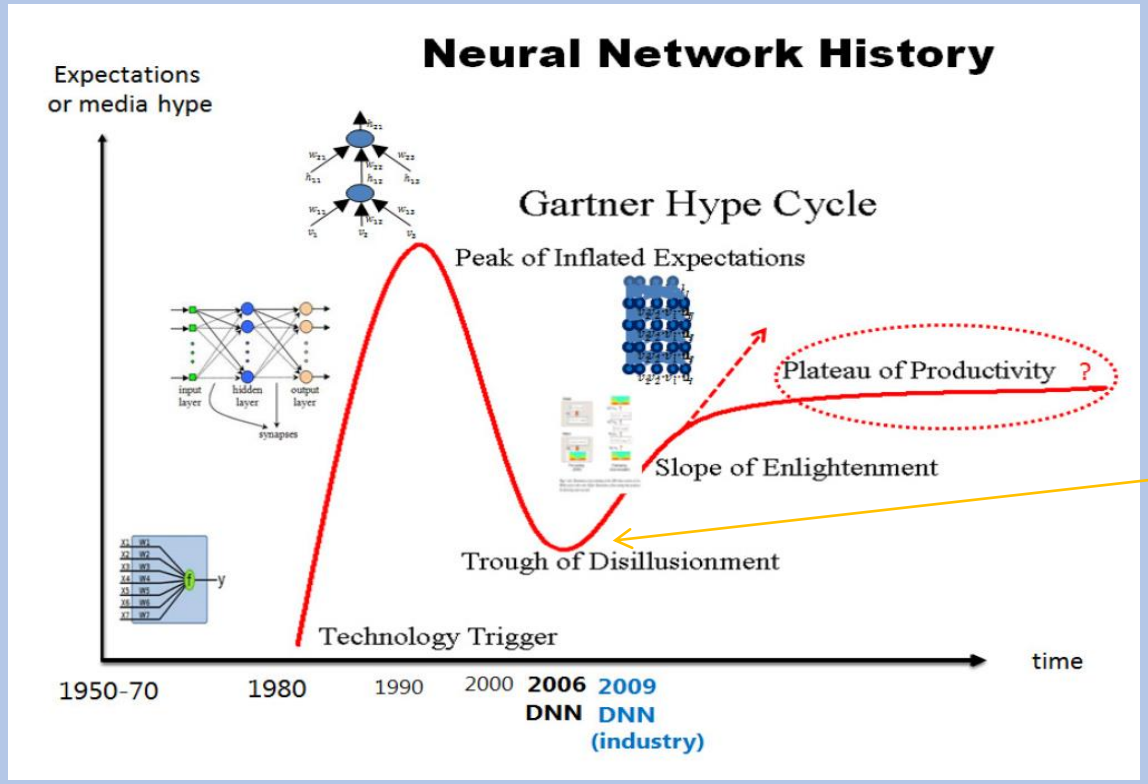
京东人工智能研究院

NLPCC 2019， 敦煌

提纲

- 自然语言进展
  - 语音和自然语言处理
- 语言+视觉多模态智能
  - 图像描述（Image Caption）
  - 视觉问答（VQA）
  - 基于文字描述合成图像（Text-to-image Synthesis）
- 产业界应用

深度学习发展的盖特纳曲线图



Geoff Hinton

Yoshua Bengio

Yann LeCun

从学术界到工业界：Scale up！

NIPS08 workshop
我们邀请Hinton来做讲座，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

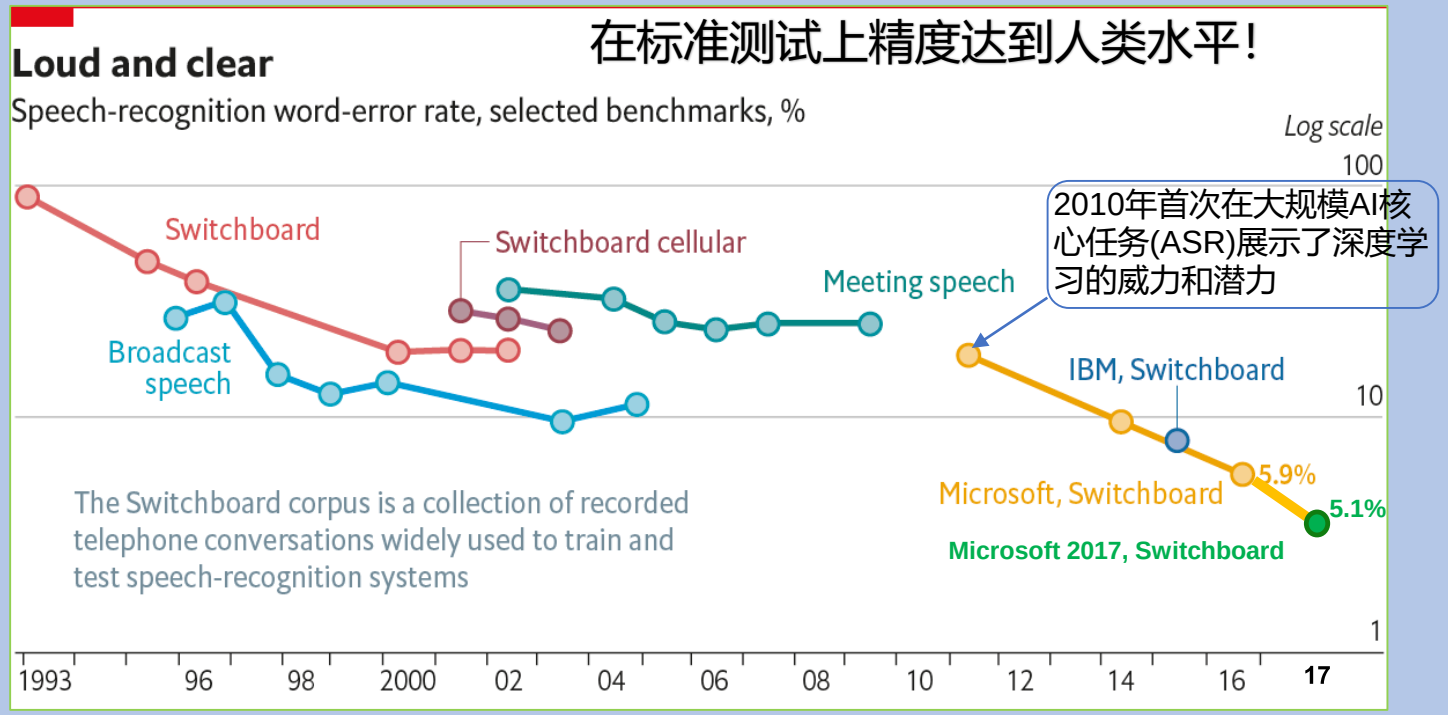
INVITED TALKS

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

语音识别


在标准测试上精度达到人类水平！
Loud and clear
Speech-recognition word-error rate, selected benchmarks, %
Log scale
100
10
1
Switchboard
Switchboard cellular
Meeting speech
Broadcast speech
2010年首次在大规模AI核心任务(ASR)展示了深度学习的威力和潜力
IBM, Switchboard
5.9%
Microsoft, Switchboard
5.1%
Microsoft 2017, Switchboard
The Switchboard corpus is a collection of recorded telephone conversations widely used to train and test speech-recognition systems
1993
96
98
2000
02
04
06
08
10
12
14
16
17

同声翻译

The universal translator on “Star Trek” comes true…

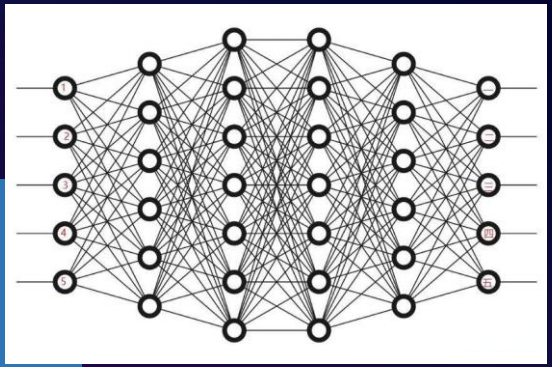
The New York Times

Scientists See Promise in Deep-Learning Programs (John Markoff November 23, 2012)

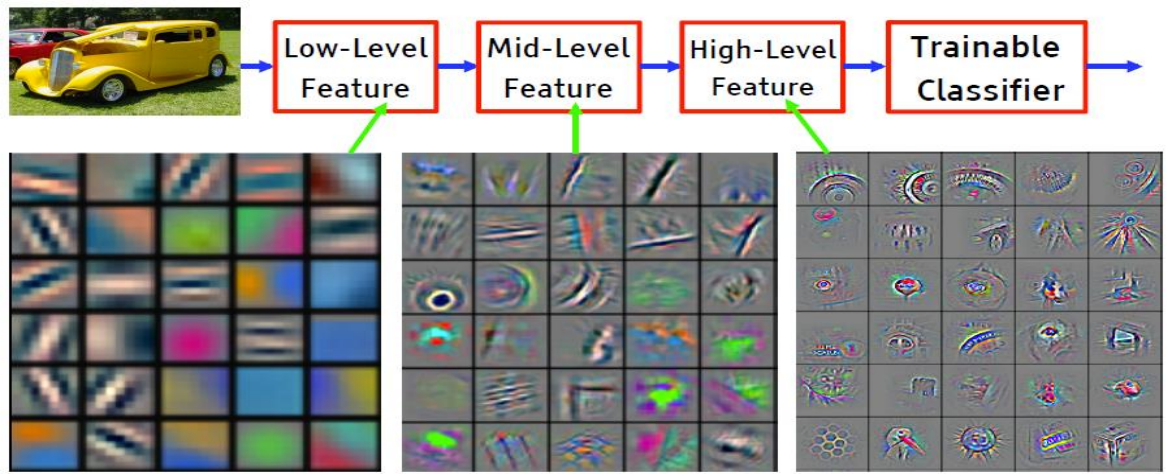
Rick Rashid in **Tianjin, China**, October, 25, 2012

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft’s top scientist, into Chinese.

为何要深度学习?



- Hierarchy of representations with increasing level of abstraction
- Each layer is a trainable feature transform
- 图像: 像素 → 边缘检测 → 纹路 → 形状 → 部位 → 物体
- 文字: 边旁部首 → 字 → 词 → 短语 → 句子 → 长篇故事



自然语言处理

- 让AI理解人类
  - 理解意图，解析语义，识别情绪，搜索推荐
- 让AI能被人类理解
  - 文本摘要，内容生成，话题展开，情感对话

文本分类、搜索、推荐、广告、等应用

分类、检索

阅读理解、问答、知识图谱构建等

语义解析

机器写作、文本摘要、文案生成、等

内容创作

情感客服、零售导购、咨询交谈、语义交互音箱/IOT等

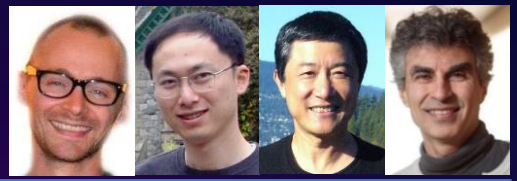
人机多轮对话

语言理解/语义槽值提取

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

2013年首次成功
应用RNN for SLU



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

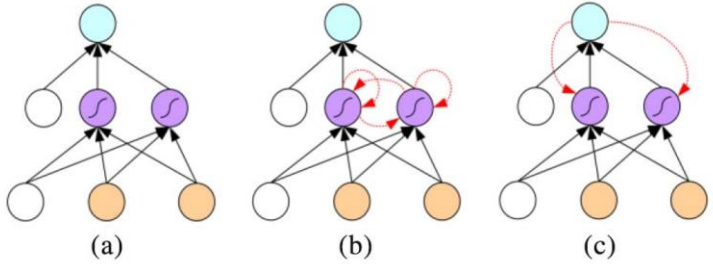


Fig. 1. Three types neural networks. (a) Feed-forward NN; (b) Elman-RNN; (c) Jordan-RNN.

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

【Mesnil, He, Deng, Bengio, InterSpeech2013】

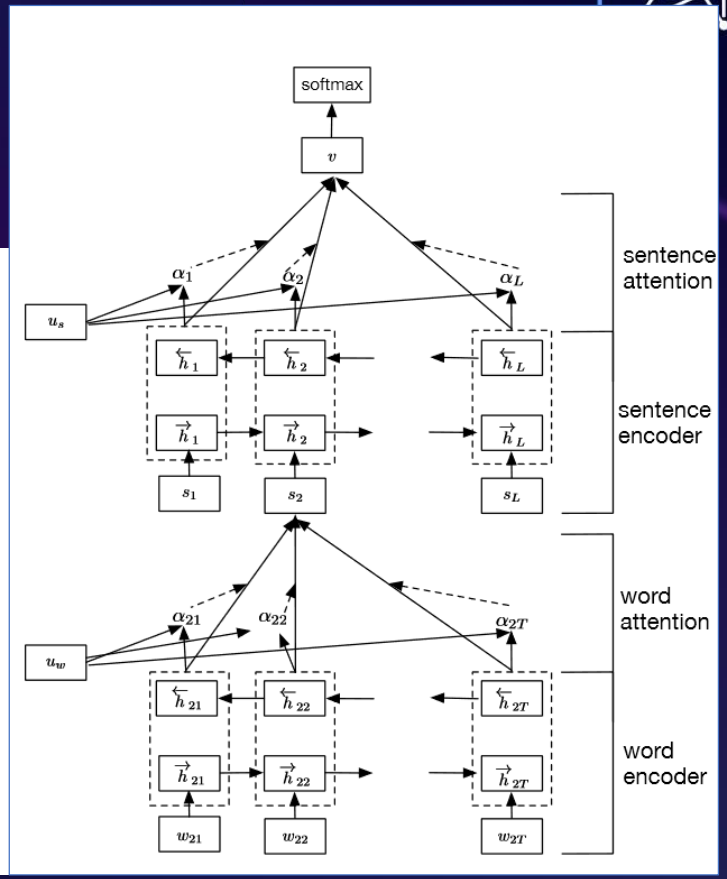
语言理解/意图分类

Hierarchical Attention Net (HAN)

我们于2016年提出层次化注意力模型 (HAN)，以更好地在词、句子、段落、等多个层面来理解语言，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



【Yang, Yang, Dyer, He, Smola, Hovy, NAACL2016】

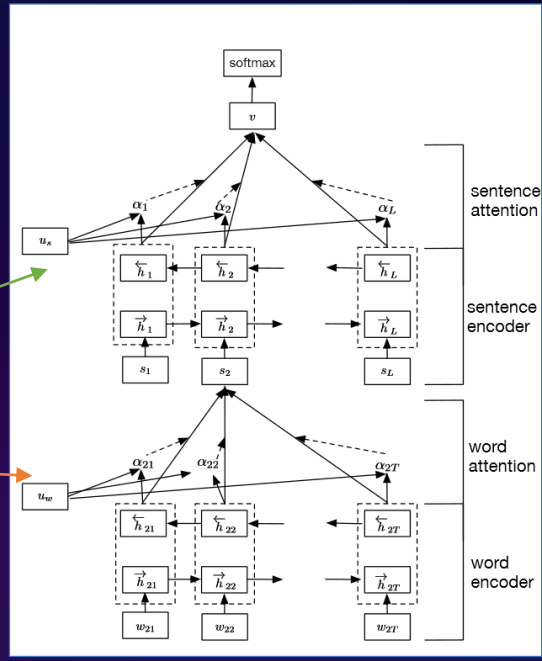
层次化注意力网络 (HAN)

Word Attention： weight more on informative words (w.r.t. the sentence)

$$u_{it} = \tanh(W_w h_{it} + b_w)$$
$$\alpha_{it} = \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)}$$
$$s_i = \sum_t \alpha_{it} h_{it}.$$

Sentence Attention： weight more on informative sentences (w.r.t. the document)

$$u_i = \tanh(W_s h_i + b_s),$$
$$\alpha_i = \frac{\exp(u_i^\top u_s)}{\sum_i \exp(u_i^\top u_s)},$$
$$v = \sum_i \alpha_i h_i,$$



The key roles of u_w and u_s (word and sentence level context vectors):

- note that u_w and u_s don't represent an actual word or sentence, they can be viewed as representing *a virtual concept of "importance"*)
- A self-attention mechanism that helps to better learn sentence and document representation
- Both are learned together with other parameters of the model end-to-end.

classification:

$$p = \text{softmax}(W_c v + b_c).$$

【Yang, Yang, Dyer, He, Smola, Hovy, NAACL2016】

层次化注意力网络（HAN）

	Methods	Yelp'13	Yelp'14	Yelp'15	IMDB	Yahoo Answer	Amazon
Zhang et al., 2015	BoW	-	-	58.0	-	68.9	54.4
	BoW TFIDF	-	-	59.9	-	71.0	55.3
	ngrams	-	-	56.3	-	68.5	54.3
	ngrams TFIDF	-	-	54.8	-	68.5	52.4
	Bag-of-means	-	-	52.5	-	60.5	44.1
Tang et al., 2015	Majority	35.6	36.1	36.9	17.9	-	-
	SVM + Unigrams	58.9	60.0	61.1	39.9	-	-
	SVM + Bigrams	57.6	61.6	62.4	40.9	-	-
	SVM + TextFeatures	59.8	61.8	62.4	40.5	-	-
	SVM + AverageSG	54.3	55.7	56.8	31.9	-	-
	SVM + SSWE	53.5	54.3	55.4	26.2	-	-
Zhang et al., 2015	LSTM	-	-	58.2	-	70.8	59.4
	CNN-char	-	-	62.0	-	71.2	59.6
	CNN-word	-	-	60.5	-	71.2	57.6
Tang et al., 2015	Paragraph Vector	57.7	59.2	60.5	34.1	-	-
	CNN-word	59.7	61.0	61.5	37.6	-	-
	Conv-GRNN	63.7	65.5	66.0	42.5	-	-
	LSTM-GRNN	65.1	67.1	67.6	45.3	-	-
This paper	HN-AVE	67.0	69.3	69.9	47.8	75.2	62.9
	HN-MAX	66.9	69.3	70.1	48.2	75.2	62.9
	HN-ATT	**68.2**	**70.5**	**71.0**	**49.4**	**75.8**	**63.6**

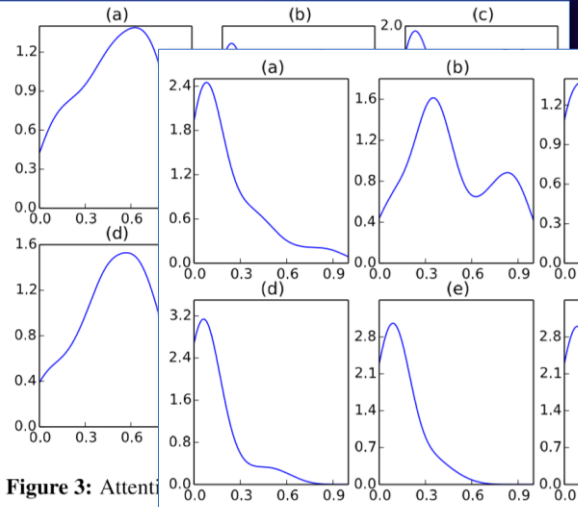


Figure 3: Attenti
gate distribution
with ratings 1-5 r
bution shifts to *hi*

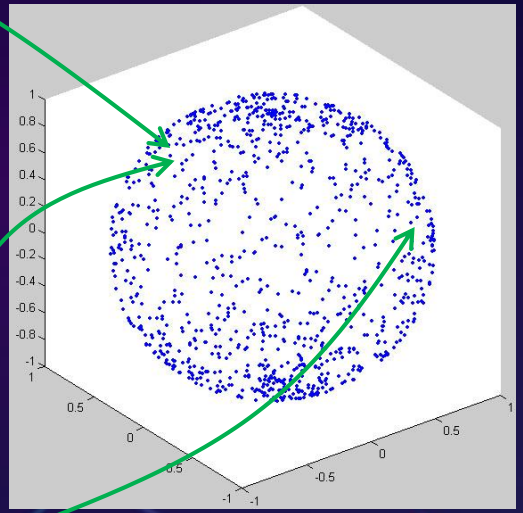
Figure 4: Attention weight distribution of the word `bad`. The setup is as above: (a) contains the aggregate distribution, while (b)-(f) contain stratifications to reviews with ratings 1-5 respectively. Contrary to before, the word `bad` is considered important for poor ratings and less so for good ones.

【Yang, Yang, Dyer, He, Smola, Hovy, NAACL2016】

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

语义空间



抽象的语义表征

神经网络

输入

通过深度神经网络逐步抽取语义上的不变性 (invariance)

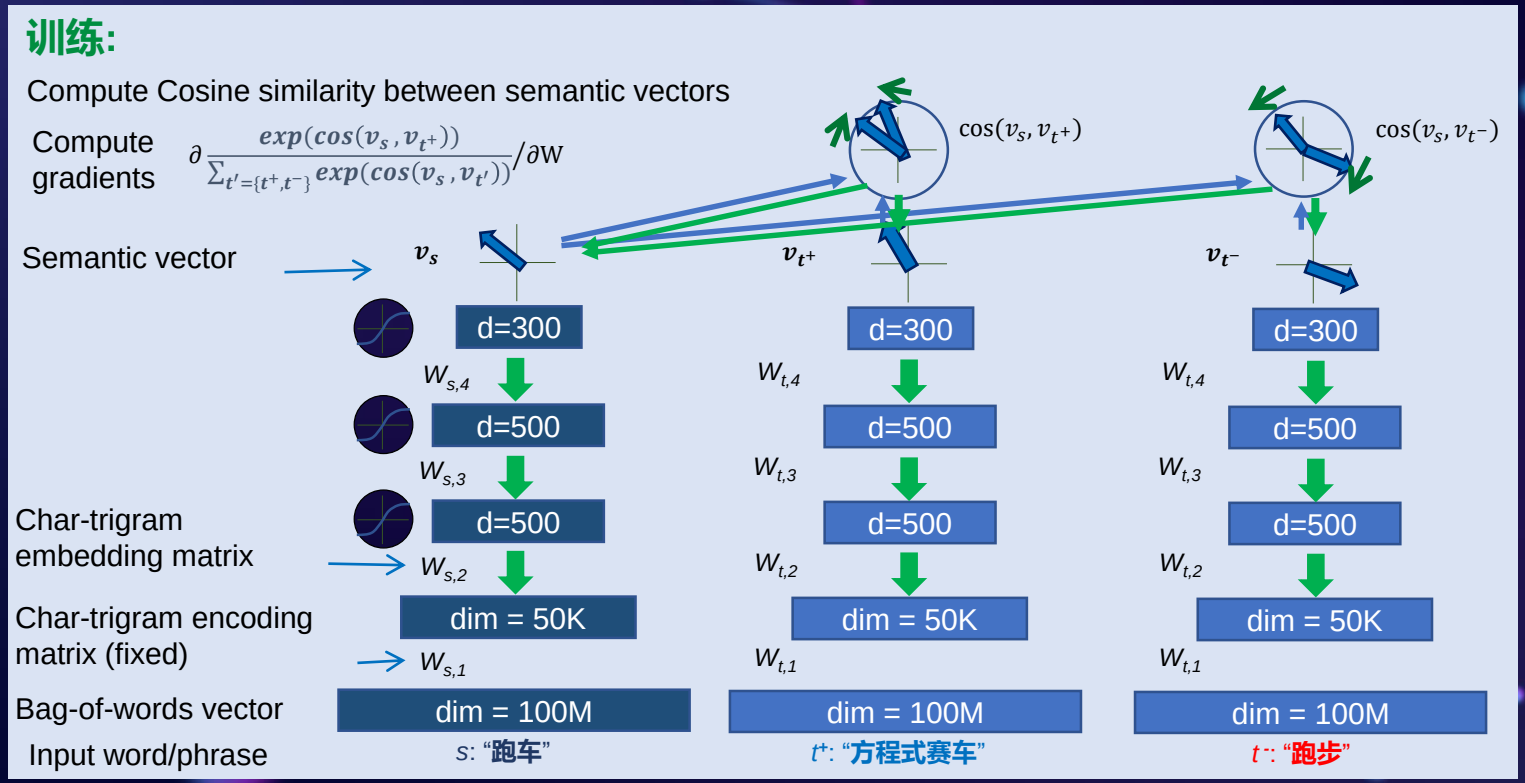
自然语言的描述 *“小明快递了一袋苹果给外公”*

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”*

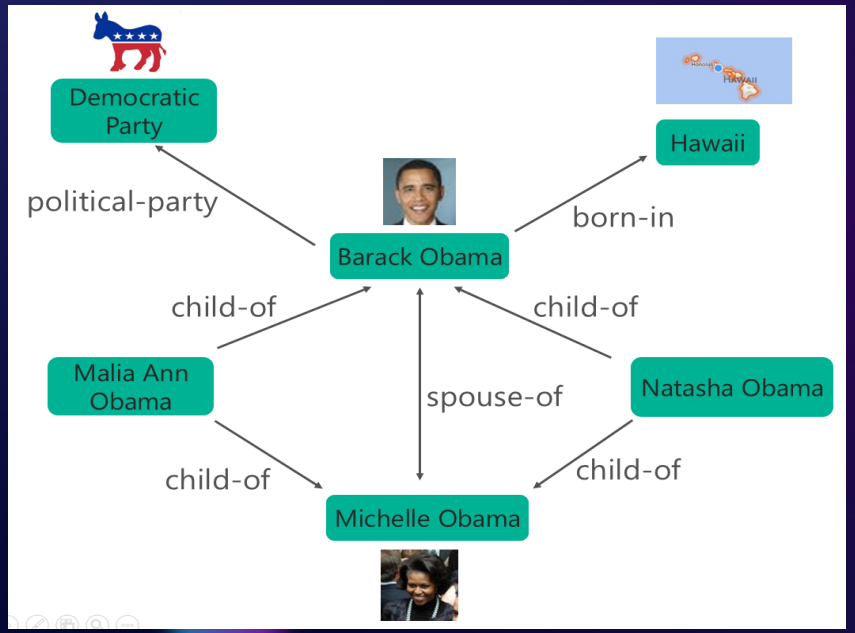
【”DSSM”, Huang, He, Gao, Deng, Acero, Heck, CIKM2013】

DSSM: 深度结构化语义模型



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

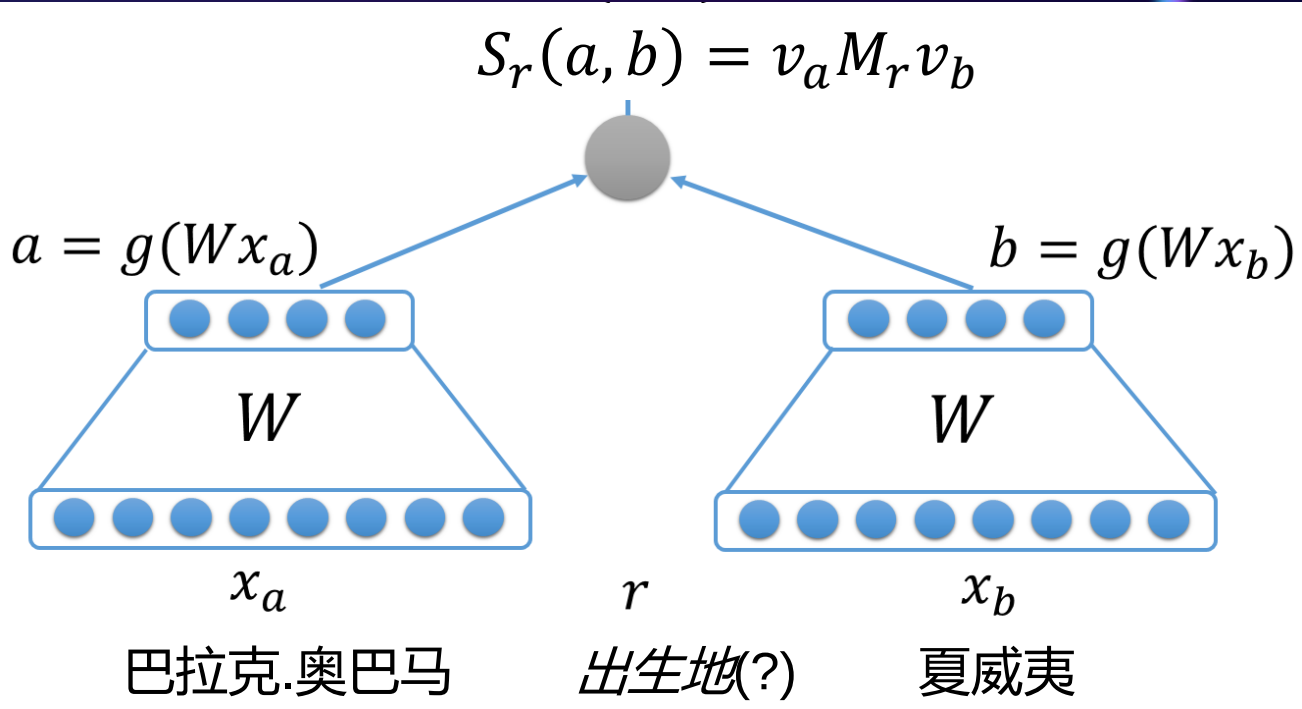
知识图谱



Freebase
DBpedia
YAGO
NELL
OpenIE/ReVerb

基于神经网络的知识表示与计算

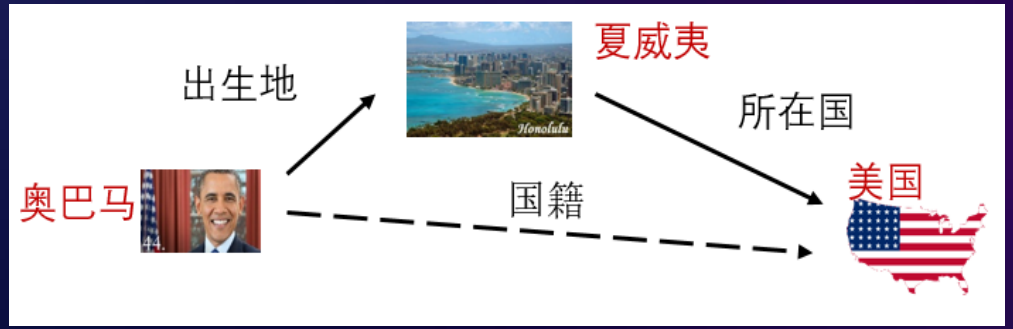
实体表示为一个向量
关系表示为一个矩阵



知识推理及知识扩展

通过在语义空间中的计算，可推理和扩展出新的知识

如：出生地$(a,b) \wedge$ 所在国$(b,c) \Rightarrow$ 国籍(a,c)



在深度语义空间，关系链可通过一系列关系矩阵来计算:

$$If \quad dist(M_r,\ M_1 \circ M_2 \circ \cdots \circ M_n) < \theta$$

$$Then\ r_1(e_1,e_2) \wedge r_2(e_2,e_3) \cdots \wedge r_n(e_n,e_{n+1}) \rightarrow r(e_1,e_{n+1})$$

【Yang, Yih, He, Gao, Deng, ICLR2015】

知识推理及问答

谁是贾斯汀.比伯的姐姐?

贾木尼.比伯

语义解析

$\lambda x.\,\text{sister_of}(\text{justin_bieber}, x)$

搜索匹配

$\text{sibling_of}(\text{justin_bieber}, x) \wedge \text{gender}(x, \text{female})$

SQL

数据库

【Yih, He, Meek, ACL2014; Yih, Chang, He, Gao, ACL2015; Golub & He EMNLP2016;…】

机器阅读理解(MRC)

机器阅读文本，回答问题

文章：印尼雅加达（CNN） - 据报道，五名欧洲人周六在印度尼西亚潜水旅行出错后获救，他们不得不在躲开科莫多巨龙的同时等待被发现。 星期四失踪后，该团体在科莫多国家公园附近的Rinca岛上的Mantaolan被发现。一同前往的潜水员向导报告共有三名英国人，一名法国人和一名瑞典人在离岛上度过了两个晚上，而这里是大型科莫多巨龙的所在地。巡逻队之后发现了他们。这个团体在周六获救。

问题: 有多少英国人获救?

答案: 三位

SQuAD

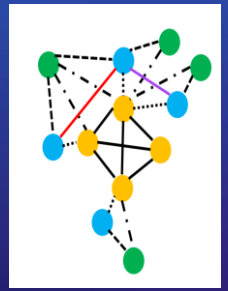
The Stanford Question Answering Dataset

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

多文档机器阅读理解

语义理解及机器问答

#	Model / Reference	Affiliation	Date	Accuracy[%]
1	JDReader (ensemble)	JD AI Research	March 2019	74.3
2	DynSAN (ensemble)	Samsung Research (SRC-B)	March 2019	73.8
3	DynSAN basic (single)	Samsung Research (SRC-B)	February 2019	71.4
4	Entity-GCN v2 (ensemble)	University of Amsterdam && University of Edinburgh	November 2018	71.2
5	HDEGraph	JD AI Research	February 2019	70.9
6	CFC	Salesforce Research	September 2018	70.6



问题：“巨魔”诞生的国家

文档1：由茱莉亚•唐纳森和大卫•罗伯茨创作的“巨魔”是一个关于巨魔和海盗的儿童故事...

文档2：茱莉亚•唐纳森是一位英文作家，剧作家和演员，同时是2011-2013年Children’s Laureate奖获得者...

文档3：Children’s Laureate奖是英国为了表彰出色的英国儿童作家或者插画家而每两年颁发一次的奖项...

答案：英国

- 跨多个文本推理
- 2019年3月，WiKiHop排名第一
- 在此数据集上超过人类表现
- 多关系多层次的异质图神经网络
- 多模态推理和关系抽取
- 可用于智能导购，内容审核，合理用药等

创作长文的技术挑战

1. 从简单输入到创作长文需要大量内容的扩充
2. 长文的内容要在大跨度上保持严谨的逻辑
3. 长文的语言结构需要反应正确的事件顺序和注意语境语态的逐步变化

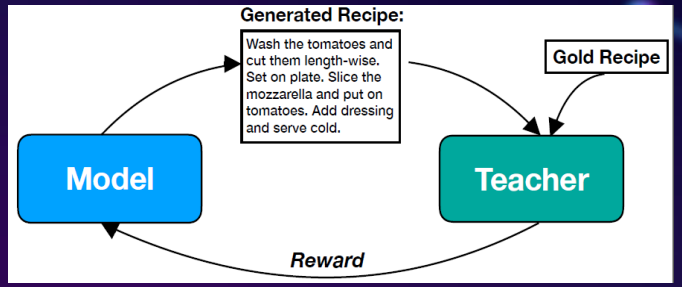
现有的端到端的模型不适合长文创作

为短文生成 (如机器翻译) 而设计, 不能抓住长文的高层语义。
度量优化, 信用分配, 维持一致性, 目标函数平衡等要从新设计。

长文创作的新进展

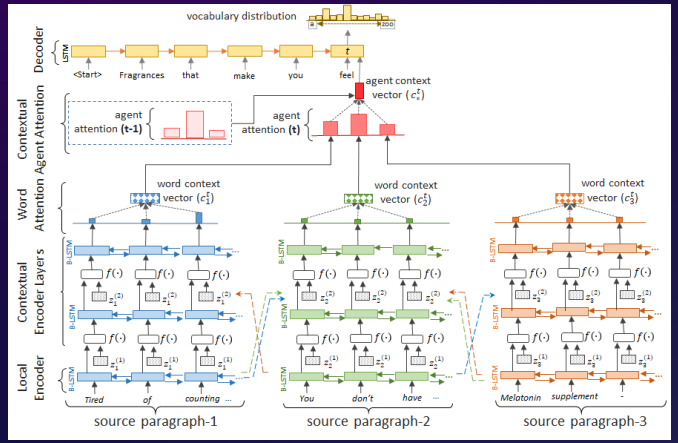
神经激励函数

捕获句子的语义属性和句子之间的连贯性及长跨度的依赖关系，从而生成逻辑顺序正确的长文本。



深度多代理通信模型

通过多个神经网络编码/译码器来突出重点概念，并保证文本连贯及语序正确。



【Celikyilmaz, Bosselut, He, Choi, NAACL2018】
【Bosselut, Celikyilmaz, He, Gao, Huang, Choi，NAACL2018】

自动生成丰富的语言内容

比如：给了原料，AI创作新菜谱

原材料及菜名

菜名:
烤草莓派

原料:
新鲜的草莓；砂糖；盐；玉米淀粉；柠檬汁；

AI生成的菜谱

1. 把糖、玉米淀粉和盐放在平底锅里。
2. 用中火加热，不断搅拌，直到混合物沸腾变浓。
3. 加入柠檬汁，加入草莓，拌匀。
4. 倒入烤馅饼的模具。
5. 在预热的325度烤箱中烤10分钟。
6. 取出，加上奶油。

【Bosselut, Celikyilmaz, He, Gao, Huang, Choi，NAACL2018】

生成诗歌

《致咿咿》

— JD AI

婴儿的眼睛，
走在爱情的森林里。
轻轻的叮咛，
轻柔的轻柔。
心中的旋律，
温柔的笑脸。

【Cheng, et al., 2018】

GPT-2 超大规模语言模型

SYSTEM PROMPT (HUMAN-WRITTEN)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

【OpenAI】

长文创作的前沿探索

比如：顶层设计与规划

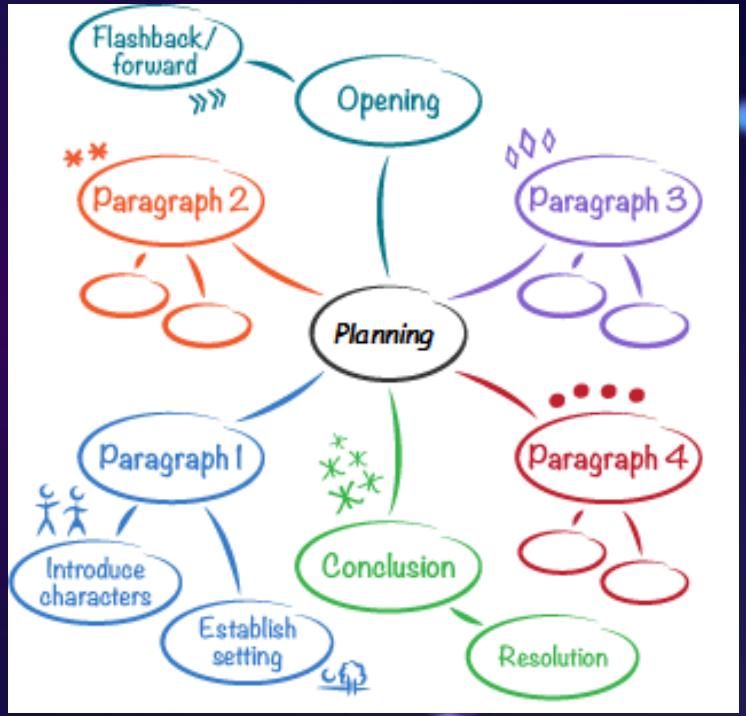
现有的文本生成模型缺乏“规划”

应先产生初略的高层主题规划

然后再对主题和子主题展开成文

最近的一些工作

多层增强学习模型及其在主题设计和长文生成中的应用



【Huang, Gan, Celikyilmaz, Wu, Wang, He, 2018】

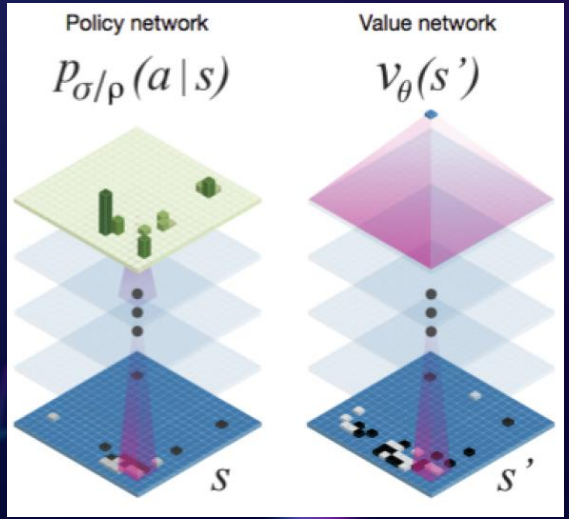
自然语言处理中的强化学习

- 人机对话, 文本游戏…
- Sequential decision tasks, 每一步:
  - 玩家阅读当前状态
  - 玩家观察可能的行动
  - 玩家需要采取正确行动，使得长期的奖励最大化
  - 玩家采取行动后，系统运行到下一步

【Narasimhan, Kulkarni, Barzilay. EMNLP2015】
【He, Chen, He, Gao, Li, Deng, Ostendorf, ACL2016】

NLP任务有非常大的执行空间

AlphaGo: 虽然围棋的观察空间非常大, 但围棋的执行空间有限



而语言的观察空间和执行空间基本是无限的

Well, here we are, back home again. The battered front door leads into the lobby.

The cat is out here with you, parked directly in front of the door and looking up at you expectantly.

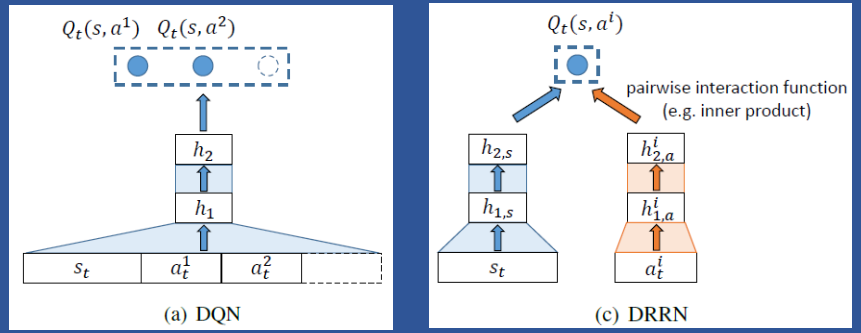
- **Step purposefully over the cat and into the lobby**
- **Return the cat's stare**
- **"Howdy, Mittens."**

Example: a snapshot of a text-based game

深度增强关联网络 DRRN

- Prior DQN work (e.g., Google DeepMind's work on Atari game, AlphaGo): state space unbounded, action space bounded.
- In NLP tasks, usually **the action space**, characterized by natural language, is discrete and nearly **unbounded**.
- We proposed a Deep Reinforcement Relevance Network **(DRRN)**
  - 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

vs.



Eval metric	Average reward		
hidden dimension	20	50	100
NN-RL (2-hidden)	0.2 (1.2)	2.6 (1.0)	3.6 (0.3)
DQN (2-hidden)	2.5 (1.3)	4.0 (0.9)	5.1 (1.1)
DRRN (2-hidden)	7.3 (0.7)	8.3 (0.7)	**10.5 (0.9)**

Results on text-based game (reward higher the better)

【He, Chen, He, Gao, Li, Deng, Ostendorf, ACL2016】

Visualizing the learning process

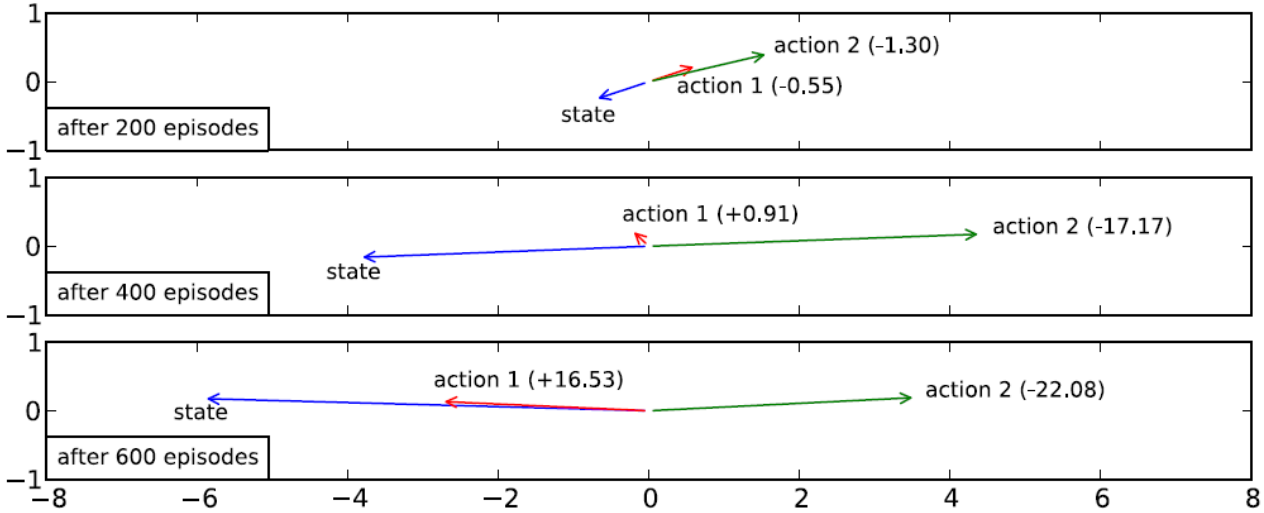
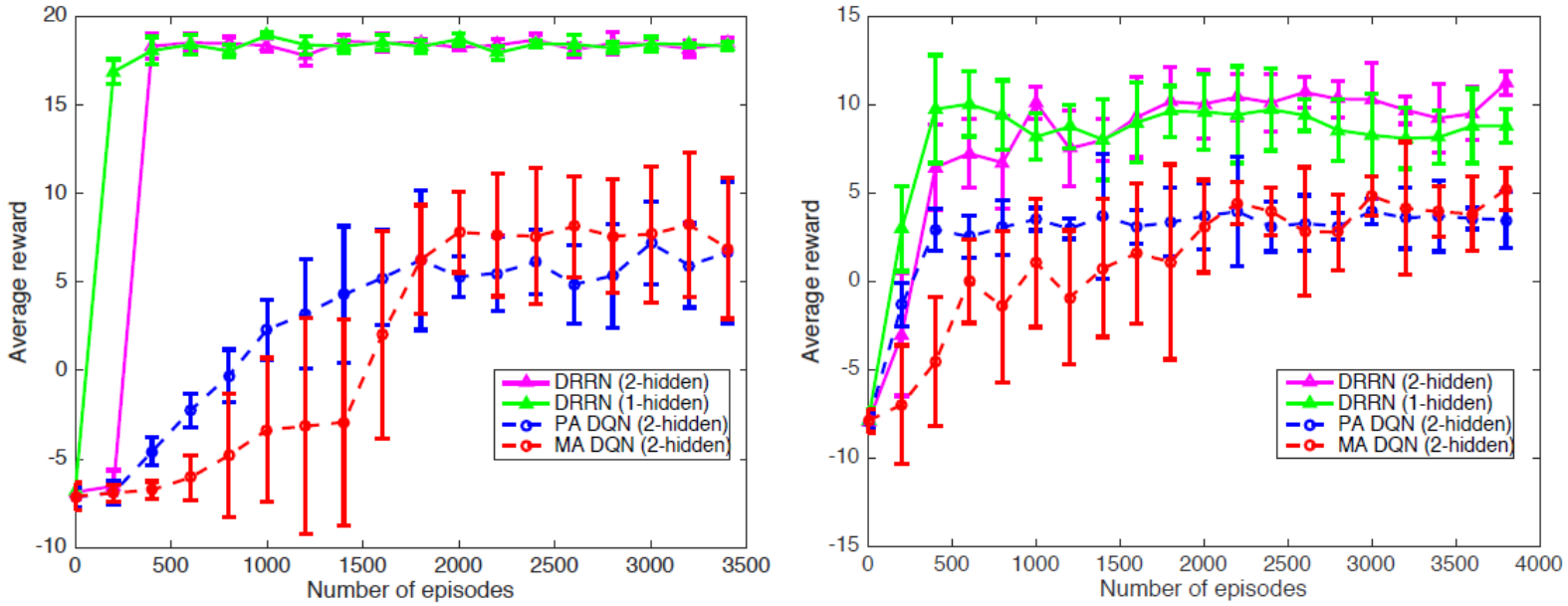


Figure 2: PCA projections of text embedding vectors for state and associated action vectors after 200, 400 and 600 training episodes. The state is "As you move forward, the people surrounding you suddenly look up with terror in their faces, and flee the street." Action 1 (good choice) is "Look up", and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

Evaluation: DRRN vs. DQN



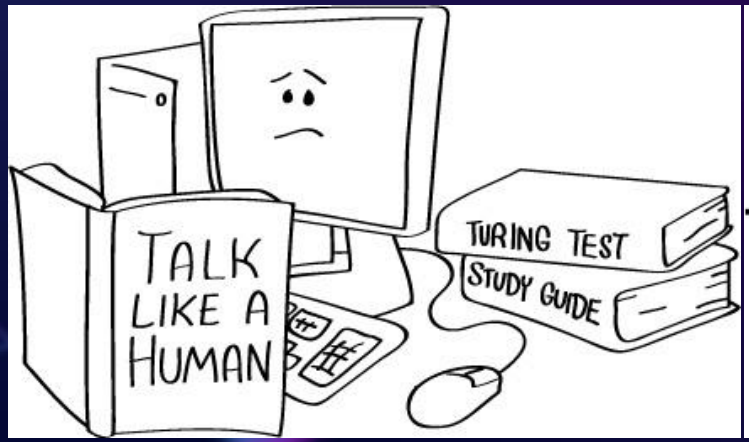
(a) Game 1: "Saving John" (b) Game 2: "Machine of Death"

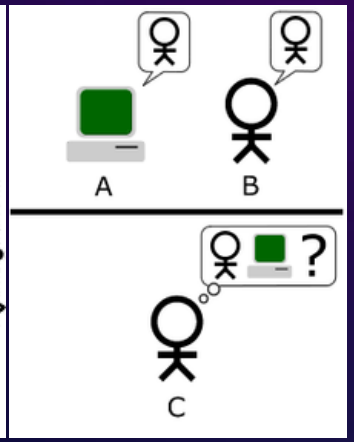
Tested on two text games

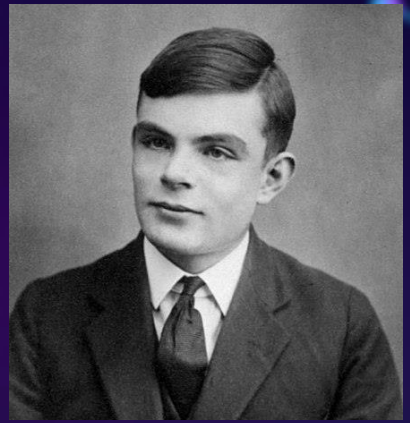
人机对话 – 人工智能皇冠上的明珠

图灵测试

通过人类和机器之间的自然语言对话来判断机器是否具有智能。







艾伦.图灵

"Computing Machinery and Intelligence" (1950)

对话系统

- 对话系统是自然语言处理技术面向终端用户应用的终极挑战之一
  - ➢对话是人类最基本最重要的交流方式
  - ➢对话也是人-机交互技术未来的期望所在
  - ➢从图灵测试提出至今，尚无系统真正通过

- 对话系统预期应用非常广
  - ➢虚拟助理：大规模复杂任务场景下帮助用户通过自然交互获取信息、解决问题、高效自动完成复杂任务。
  - ➢情感陪护：打造具备常识、同理心、正面情感个性及符合伦理的对话系统为用户提供情感交流与陪护。

对话系统发展现状：艰难探索中的蓬勃发展

京东

阿里巴巴

微软亚研

亚马逊

出现了很多特定领域对话系统的应用

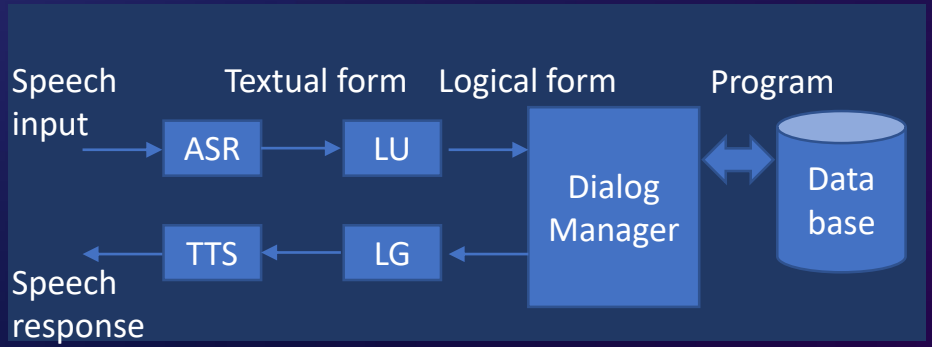
谷歌 Duplex电话系统

谷歌2018年发布的Duplex电话AI一度被认为通过了图灵测试。但尴尬的是，《纽约时报》今年5月22日的一篇报道指出，其拨打的电话中约有25%是人工打出的。而在通过自动系统打的电话中，约有15%会需要人工干预。

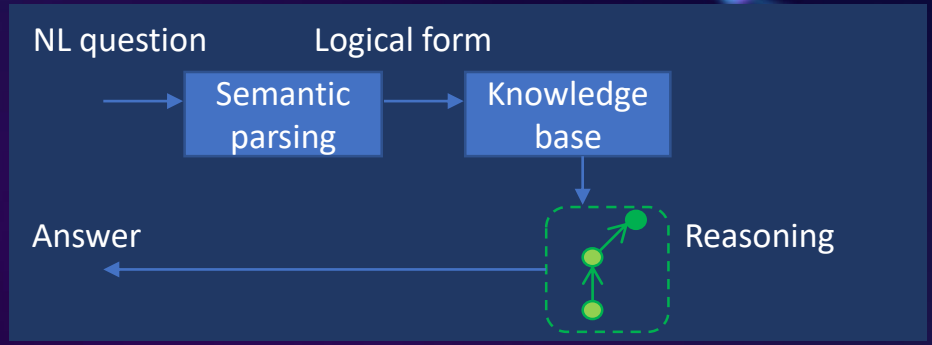
但在开放领域及针对高复杂高可靠性任务的对话系统的发展遇到瓶颈，对技术提出了重大挑战

主要的人机对话系统框架

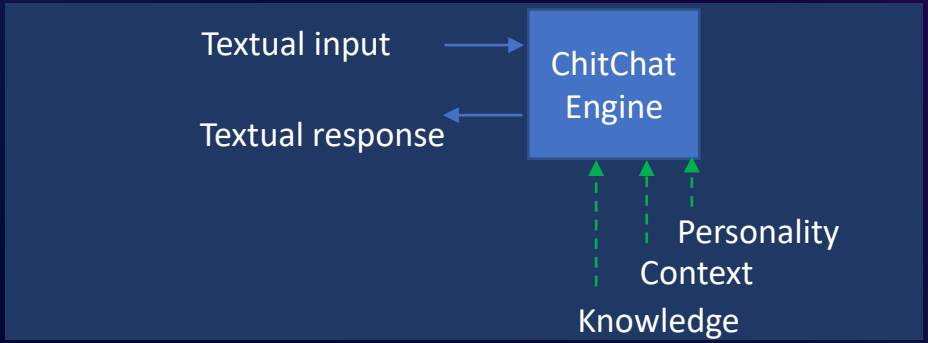
任务型对话系统：完成复杂任务



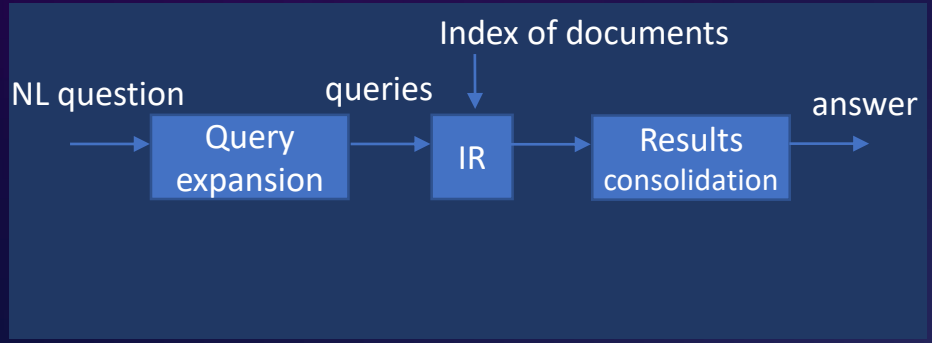
问答型对话系统: 回答问题



聊天型对话系统：情感交流与陪护



检索型对话系统：获取信息



对话机器人的机遇和挑战

我们将成为有史以来第一代与AI共生的人类

“ 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。”

— 【沈向洋, 何晓冬, 李迪, "*从Eliza到小冰：社交对话机器人的机遇和挑战*", in FITEE 2018年专刊：”人工智能 2.0：理论及应用”】

arXiv.org > cs > arXiv:1801.01957

Search or Article ID inside arXiv | All papers | Broaden your

(Help | Advanced search)

Computer Science > Artificial Intelligence

From Eliza to Xiaolce: 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(Xiaolce)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

情感智能对话

提供有“温度”的服务

- 对用户有同理心
- 有稳定可预期的个性
- 有细腻的语言技巧
- 符合社会价值观

我买的手机怎么寄的这么慢?

真不好意思，请问您是什么时候下的订单?

昨天

我查一查，系统显示已经到了。

那我怎么没收到，会不会寄丢了?

别着急，我们有保险，您放心。

我们系统显示您的邻居帮忙签收了。

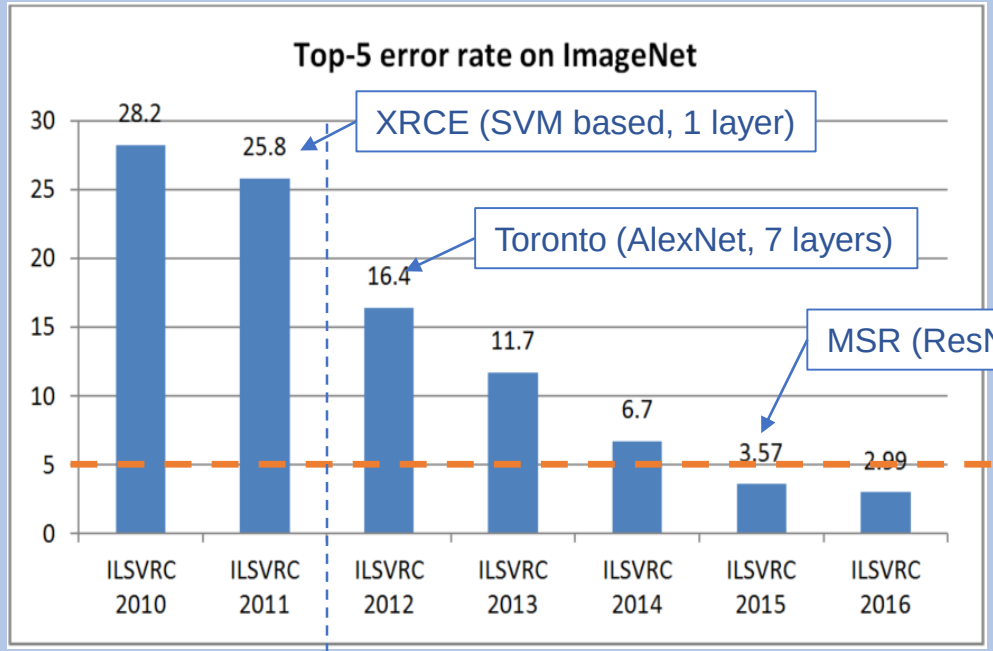
这就放心了，谢谢。

不客气。祝您愉快!

“**京东无人客服再升级：情感分析让对话型AI更有温度** – 既具备强大的闭环解决问题的能力，又兼具情感理解分析，传递情感关怀。”（http://tech.ifeng.com/）

图像识别

由于深度学习技术的进步，在ImageNet测试上AI达到人类水平



人类物体识别错误率约5%

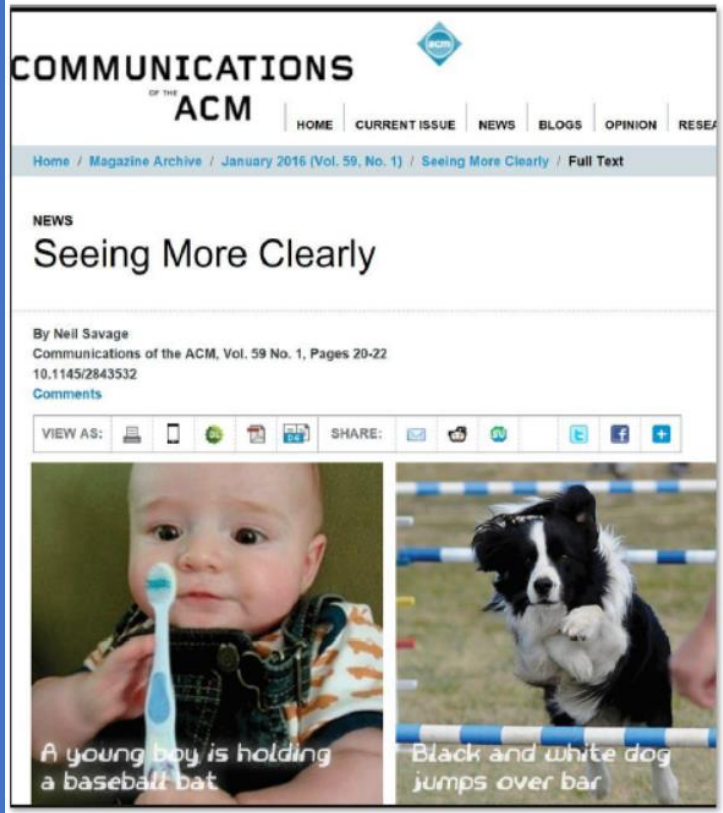
2012 前，大都是线性模型

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



"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*

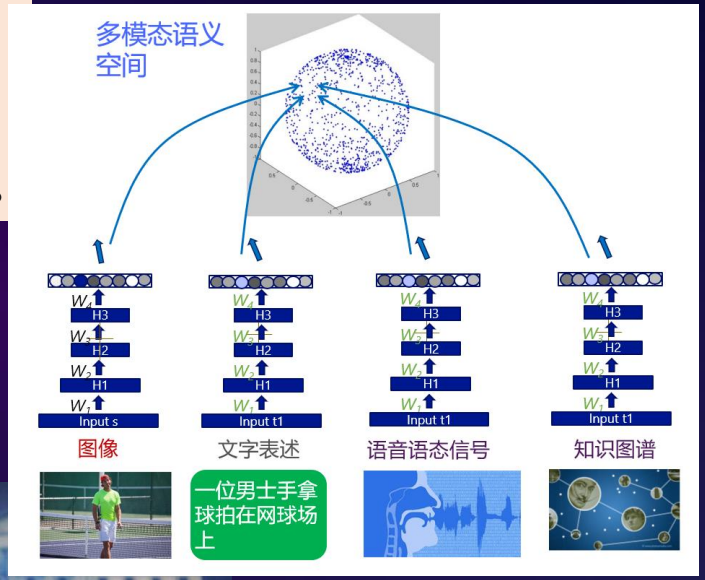
"The overall picture should have the same semantic value as the description," – *Xiaodong He*

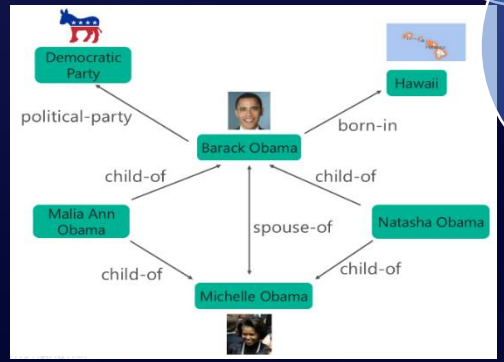
"If you really understood the image, you could answer a question about it." - *Richard Zemel*

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

多模态智能: 文本,语音,图像,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。

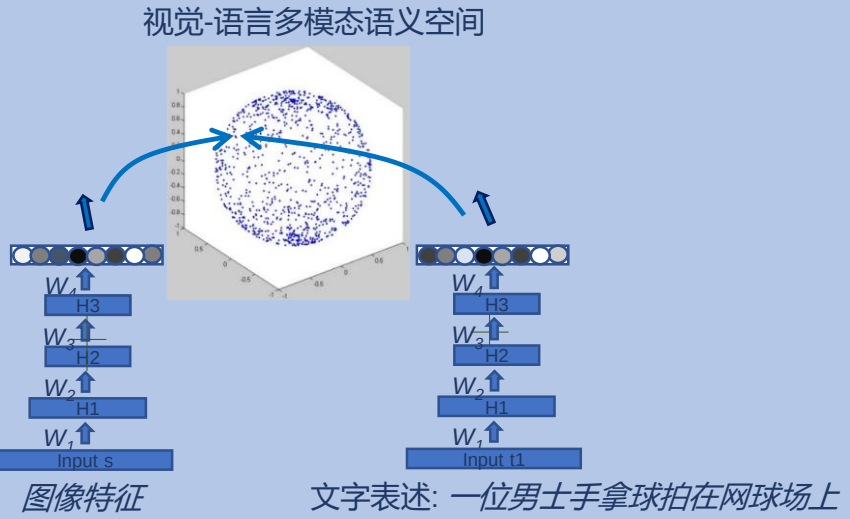




建立多模态语义空间：联结图像与文字

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



CNN

Xiaodong He, "*Deep Semantic Learning: Teach machines to understand text, image, and knowledge graph*," Keynote at DeepVision, CVPR2015

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

理解场景和知识，用语言表达

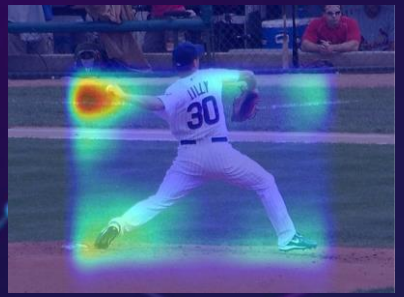
a baseball player throwing a ball

“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]

COCO 图像自动描述挑战赛 2015

COCO
Common Objects in Context

COCO Captioning Challenge 2015 @CVPR

	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, etc.			
Human	--	67.5%	63.8%

2018

加入实体知识

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.

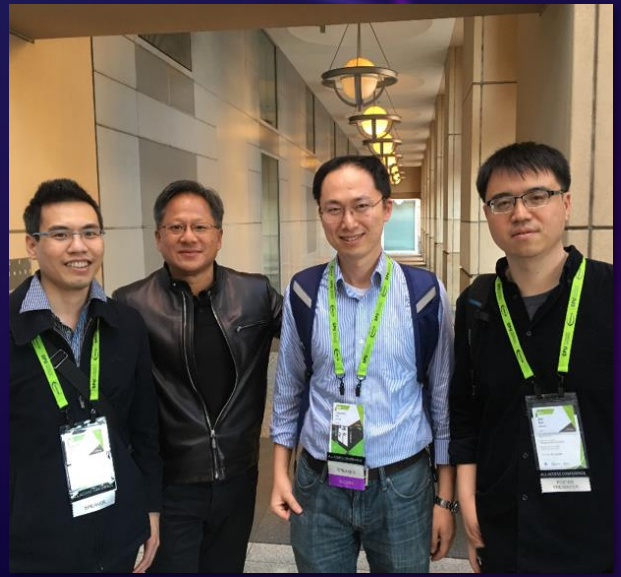
[Y Guo, L Zhang, Y Hu, X He, J Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition", ECCV 2016]

图像描述机器人： CaptionBot

[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。

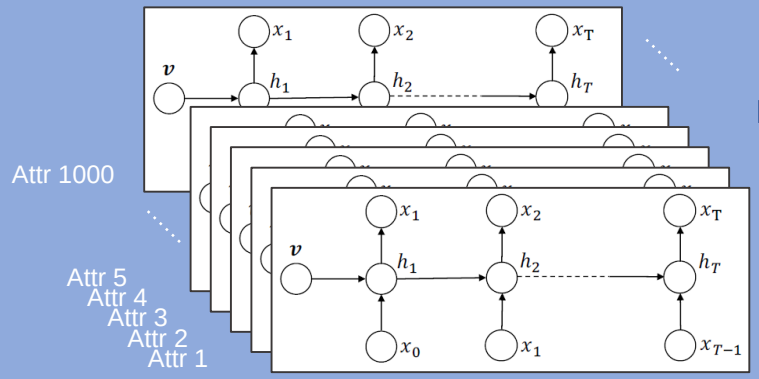
"A little boy sitting in front of a birthday cake and he seems happy." —

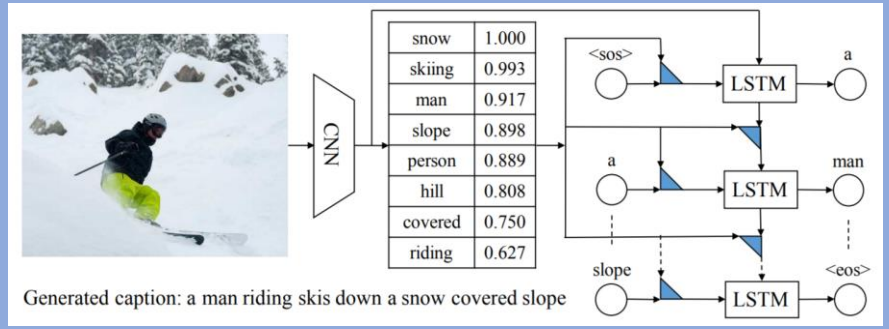
一位小男孩坐在生日蛋糕前，看起来很高兴。

语义组合网络（Semantic Compositional Net）

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

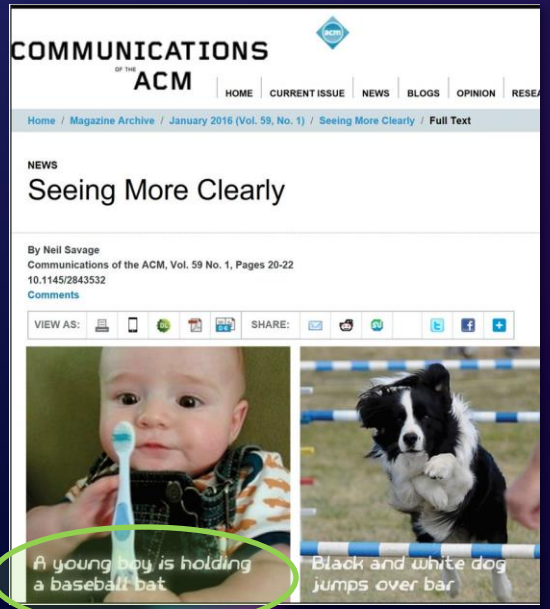




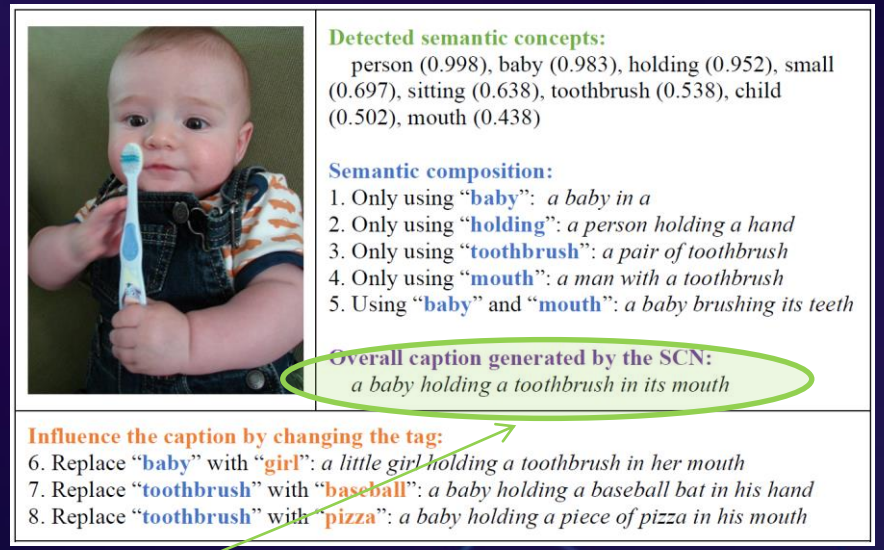
【Gan, Gan, He, Gao, Deng, CVPR17】

控制语言的语义生成

One year ago



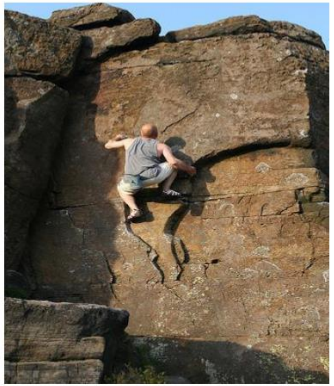
Now



【Gan, Gan, He, Gao, Deng, 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.

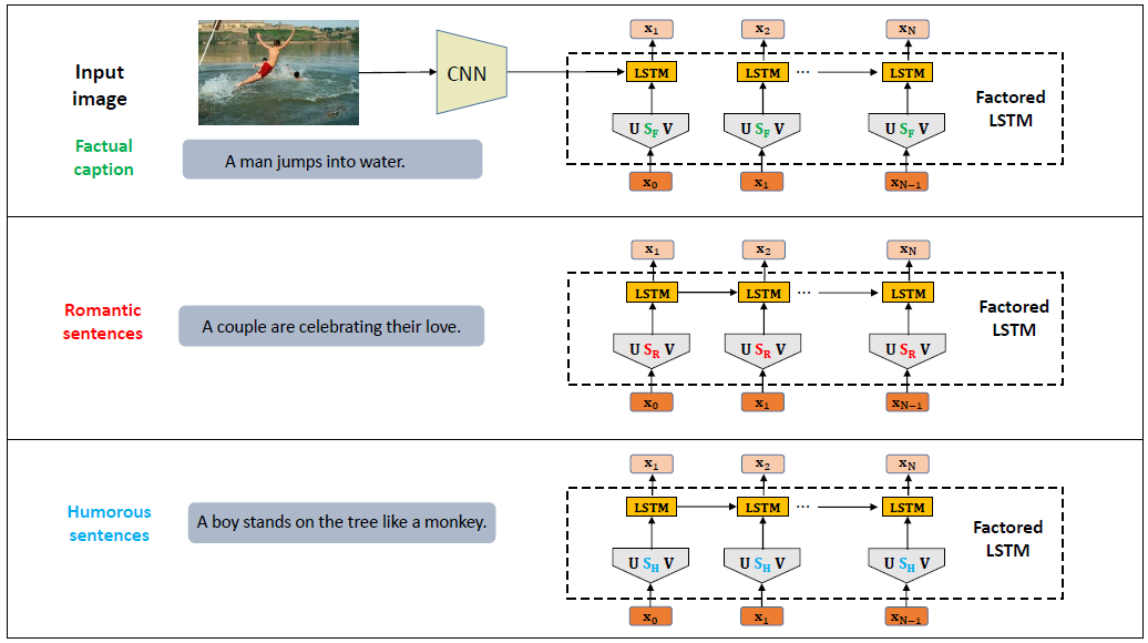
Romantic: A dog runs through the grass to meet his lover.

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

【*StyleNet*, Gan, Gan, He, Gao, Deng, CVPR2017】

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 神经网络风格模型



【*StyleNet*, Gan, Gan, He, Gao, Deng, CVPR2017】

生成带情感的语言

让AI在语言表达中加入情感，提升用户体验

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 ☺

【Huang, He, et al., 2018b】

生成语音内容

- 合成电话音频
- 评书内容合成
- 中英文混合语音合成

图文问答机器人

- 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)

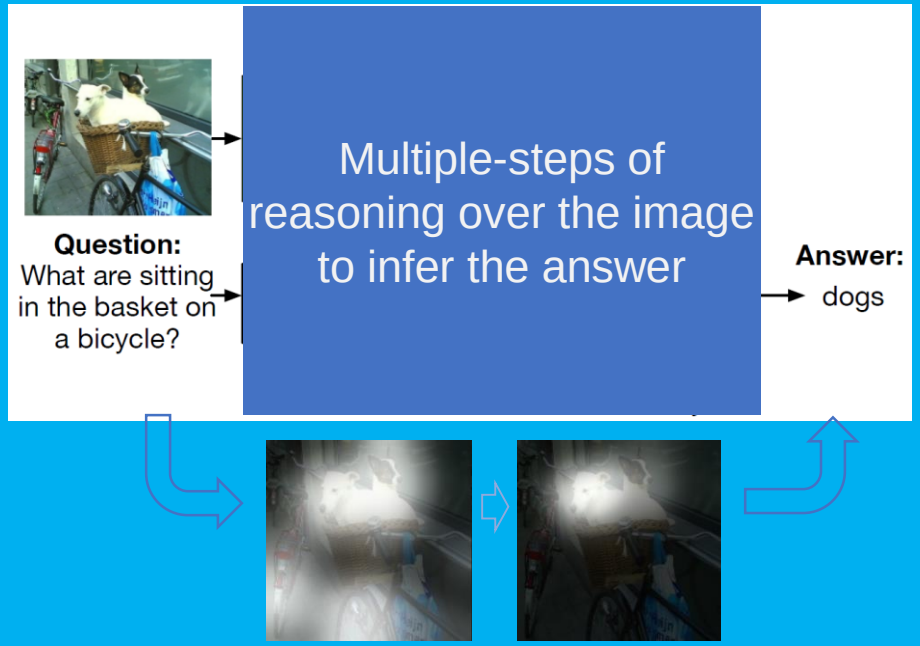
Answer:
dogs

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

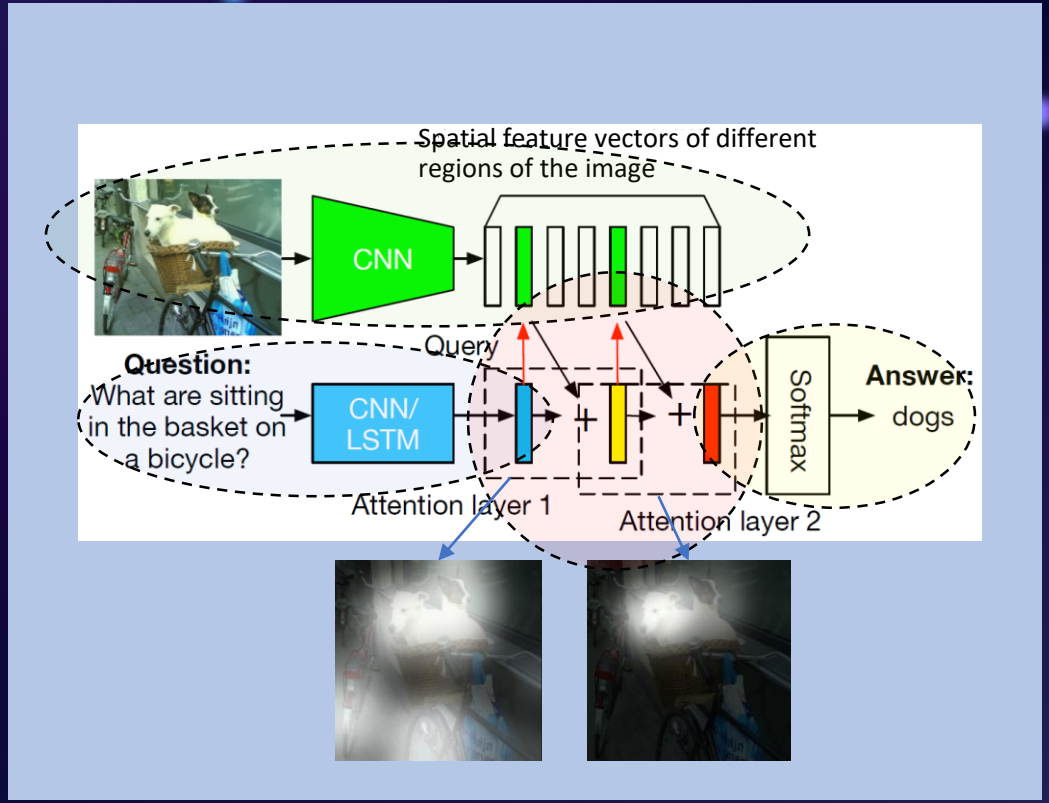
Need to focus on the specific regions that are relevant to the answer.



视觉–语言多模态推理问答

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



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

提取视觉表征

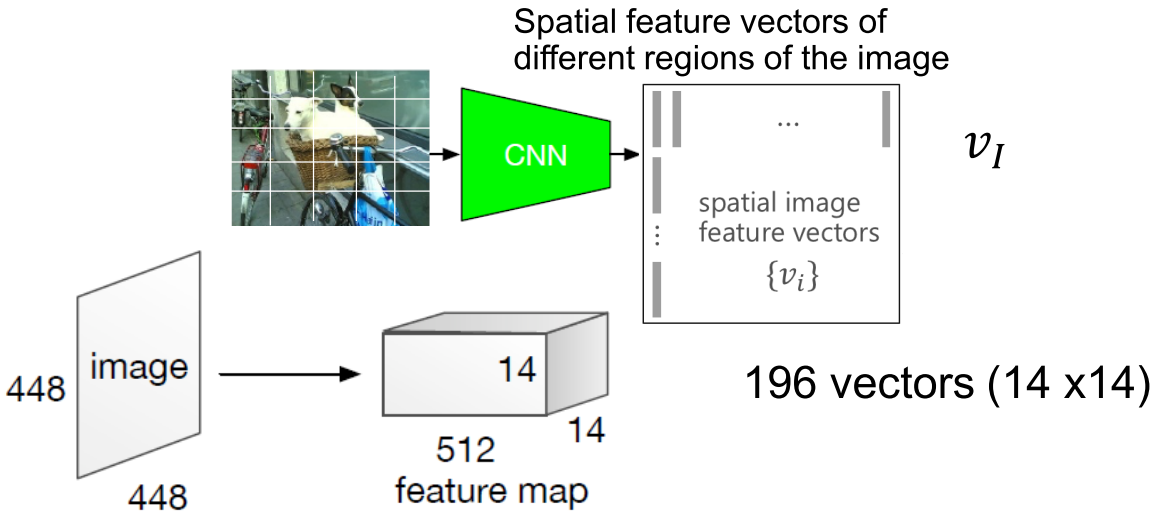
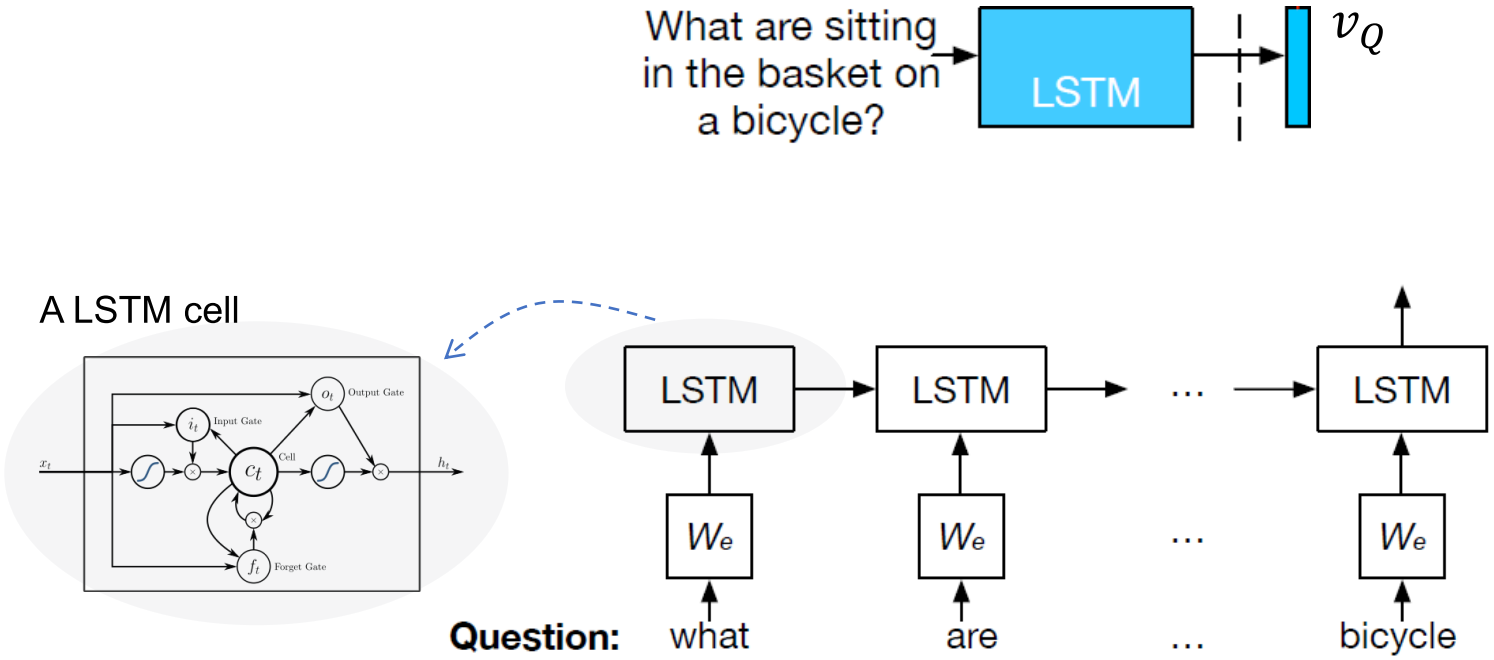


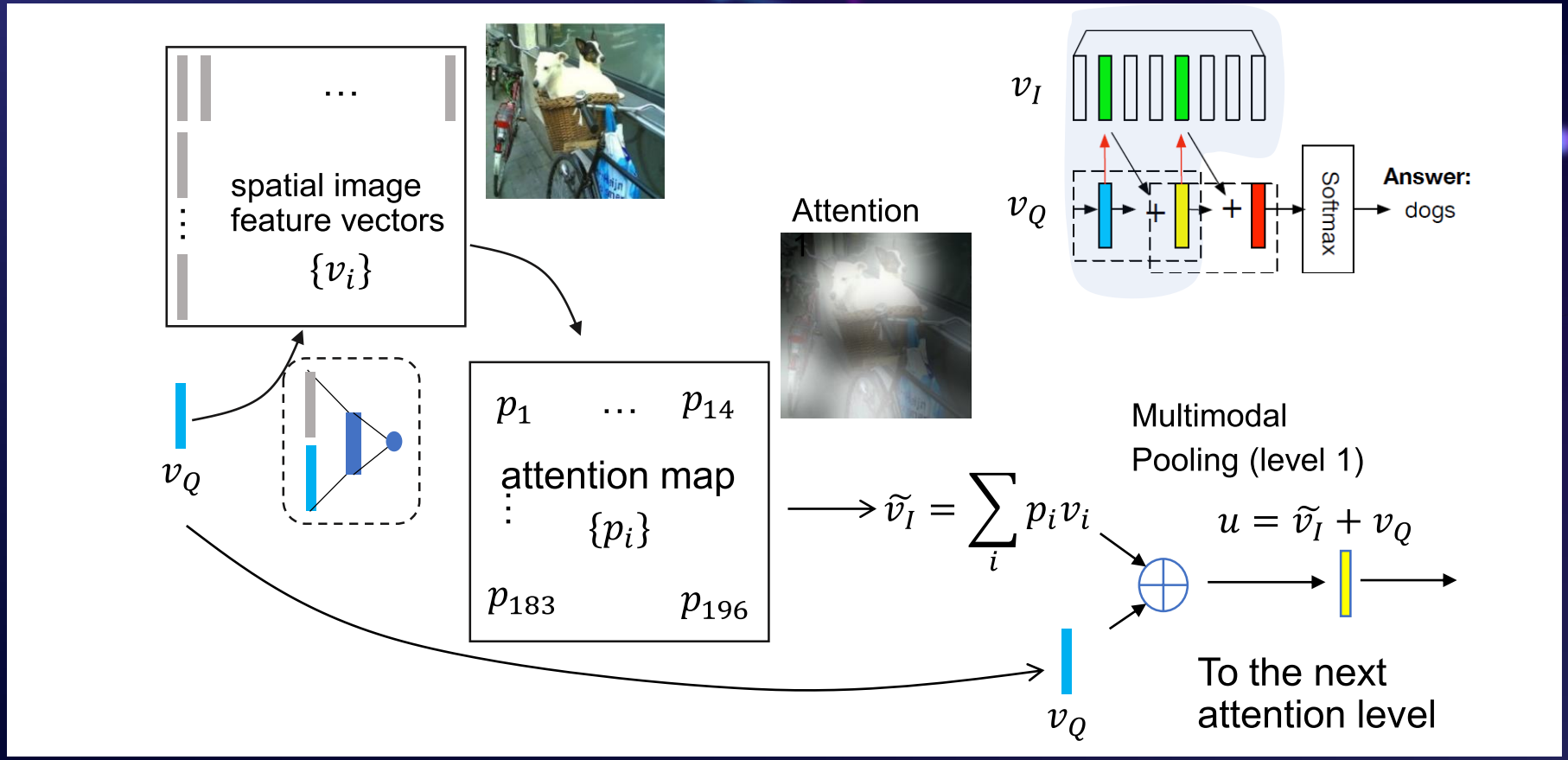
Figure 2: CNN based image model

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

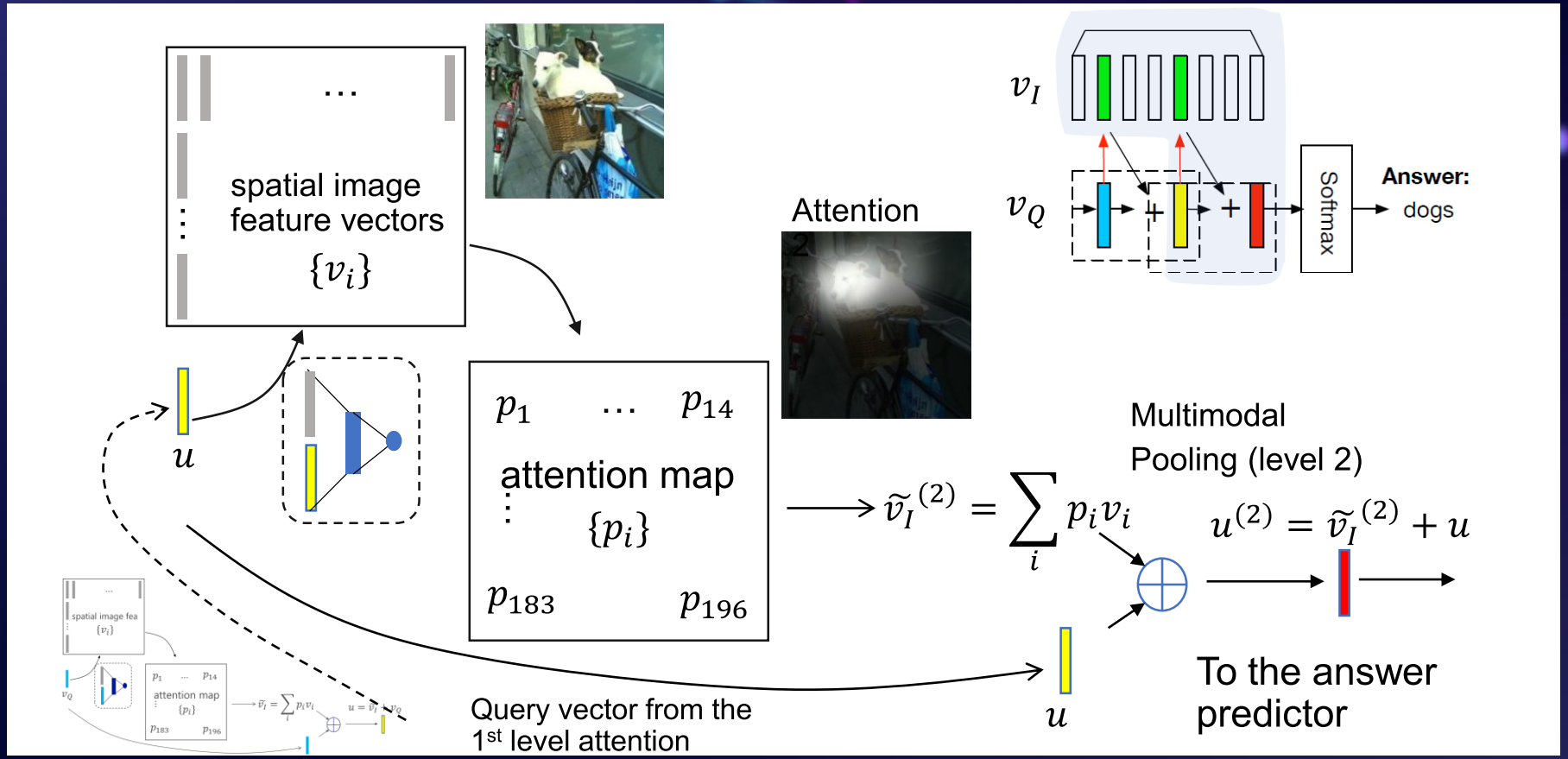
提取语言表征


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

跨模态表征融合与印证(pooling & grounding)



跨模态表征融合与印证


spatial image feature vectors
$\{v_i\}$
u
p_1 … p_{14}
attention map
$\{p_i\}$
p_{183} p_{196}
Attention
v_I
v_Q
Softmax
Answer:
dogs
Multimodal Pooling (level 2)
$\tilde{v}_I^{(2)} = \sum_i p_i v_i$
$u^{(2)} = \tilde{v}_I^{(2)} + u$
Query vector from the 1st level attention
To the answer predictor

Bottom-Up and Top-Down Attention

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

2018

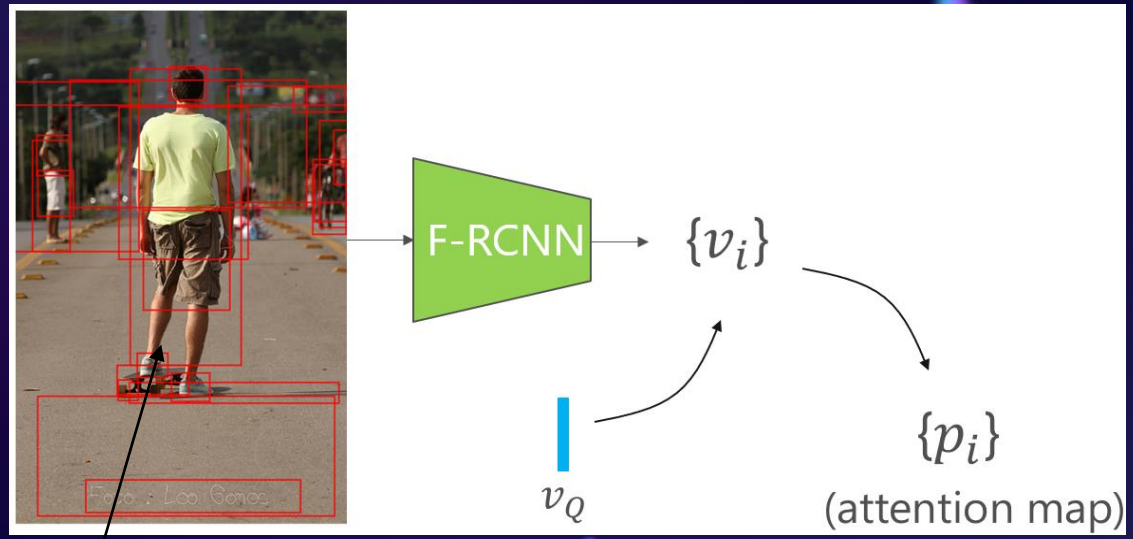
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson[1,*], **Xiaodong He**[2], **Chris Buehler**[2], **Damien Teney**[3]
Mark Johnson[4], **Stephen Gould**[1], **Lei Zhang**[2]
[1]Australian National University [2]Microsoft Research
[3]University of Adelaide [4]Macquarie University

Bottom-Up attention mechanism (new)

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

Combine Bottom-Up & Top-Down Attention

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'.

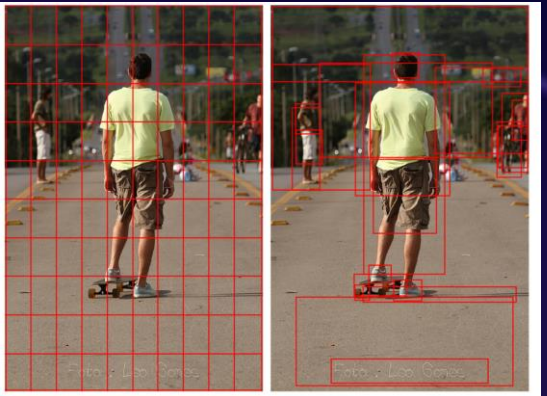
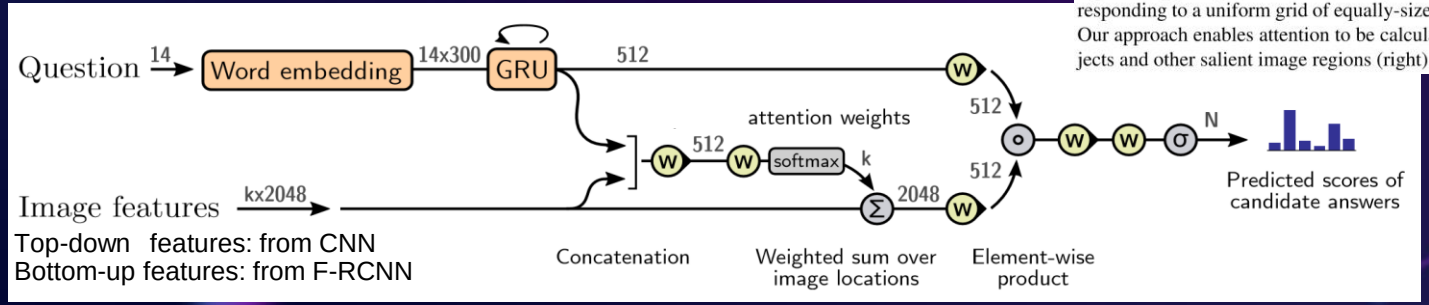


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

Overall Attention Net for VQA:



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

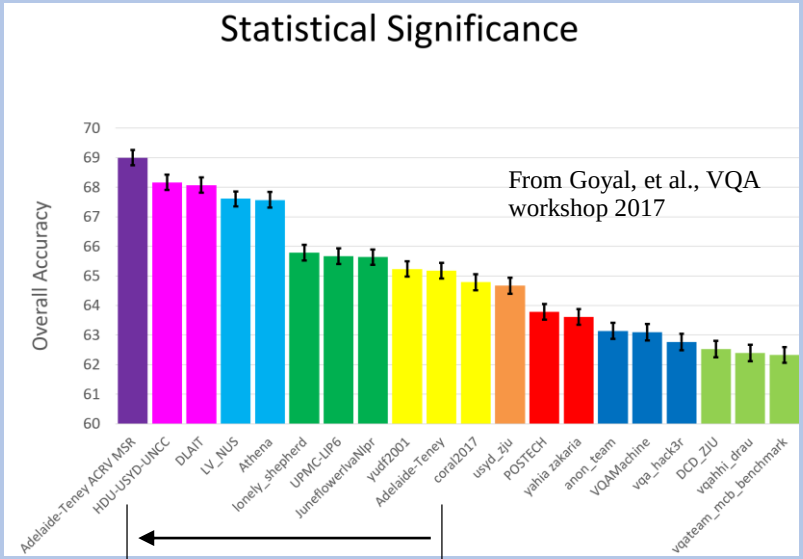
2018

Attention Example

Question: What room are they in? Answer: kitchen

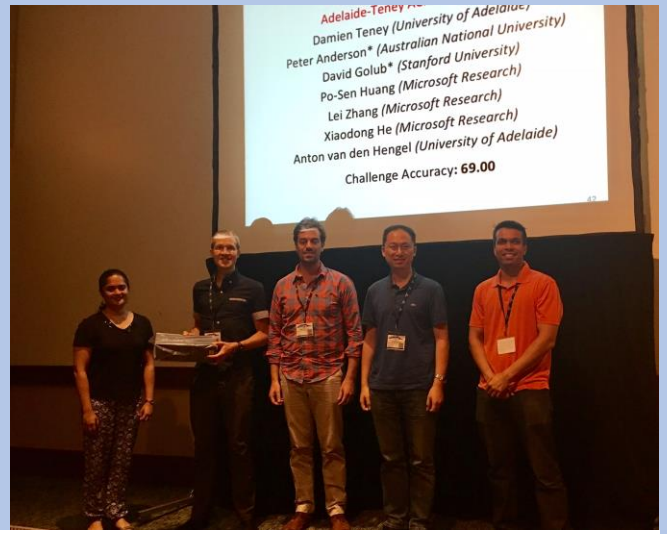
Figure 6. VQA example illustrating attention output. Given the question 'What room are they in?', the model focuses on the stove-top, generating the answer 'kitchen'.

VQA Challenge @ CVPR2017



From Goyal, et al., VQA workshop 2017

Because of **Bottom-up Attention**



[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

之后, 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。

综合图像和语言推理, 回答问题

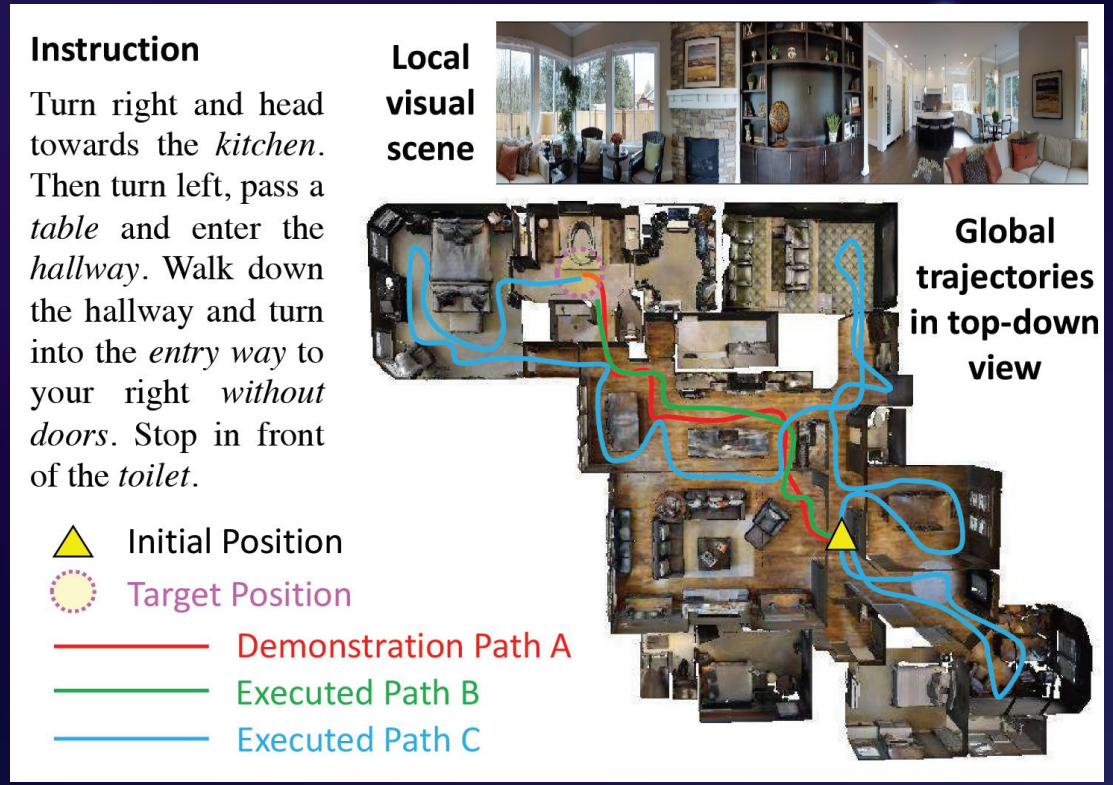
那两把蓝色椅子之间是什么?

一把伞

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

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



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

Text-to-Image generation with GANs

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

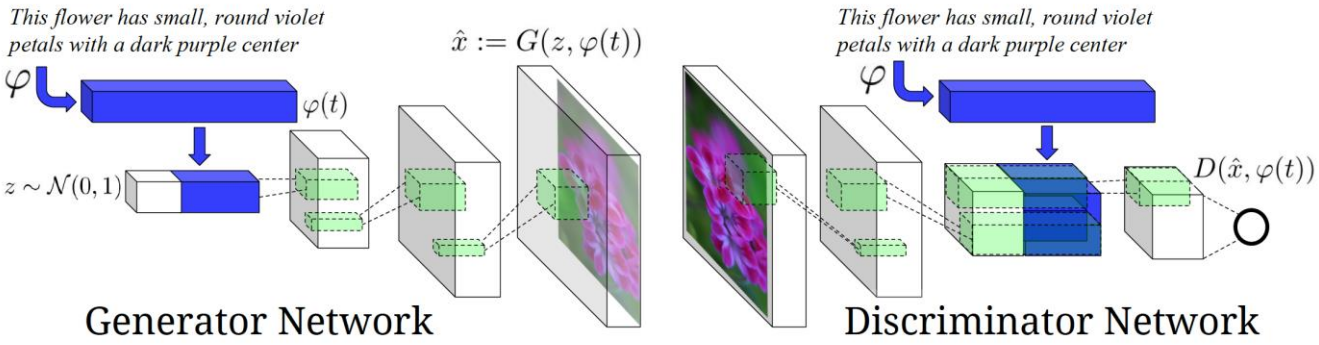


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

Objective function:

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

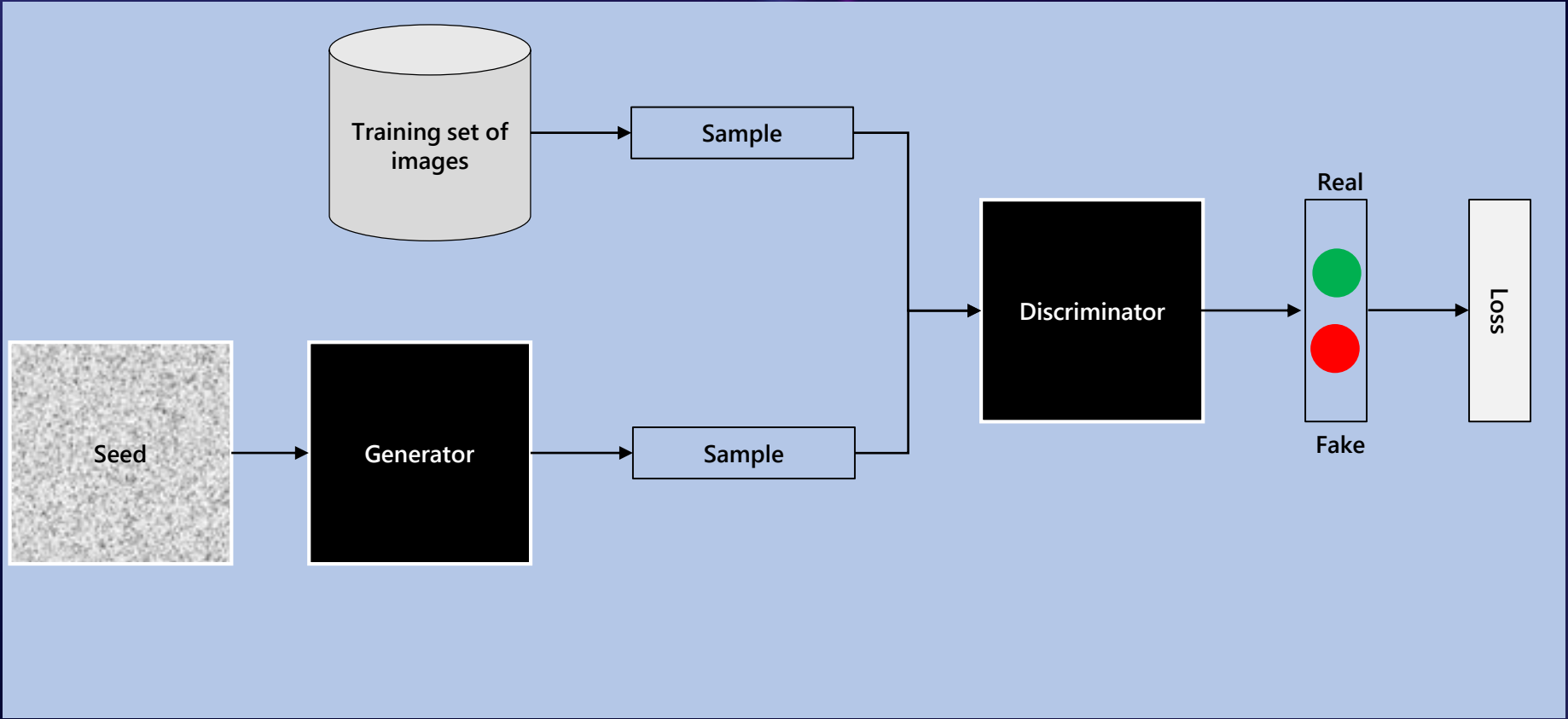
[Reed et al., Generative adversarial text-to-image synthesis, ICML, 2016]

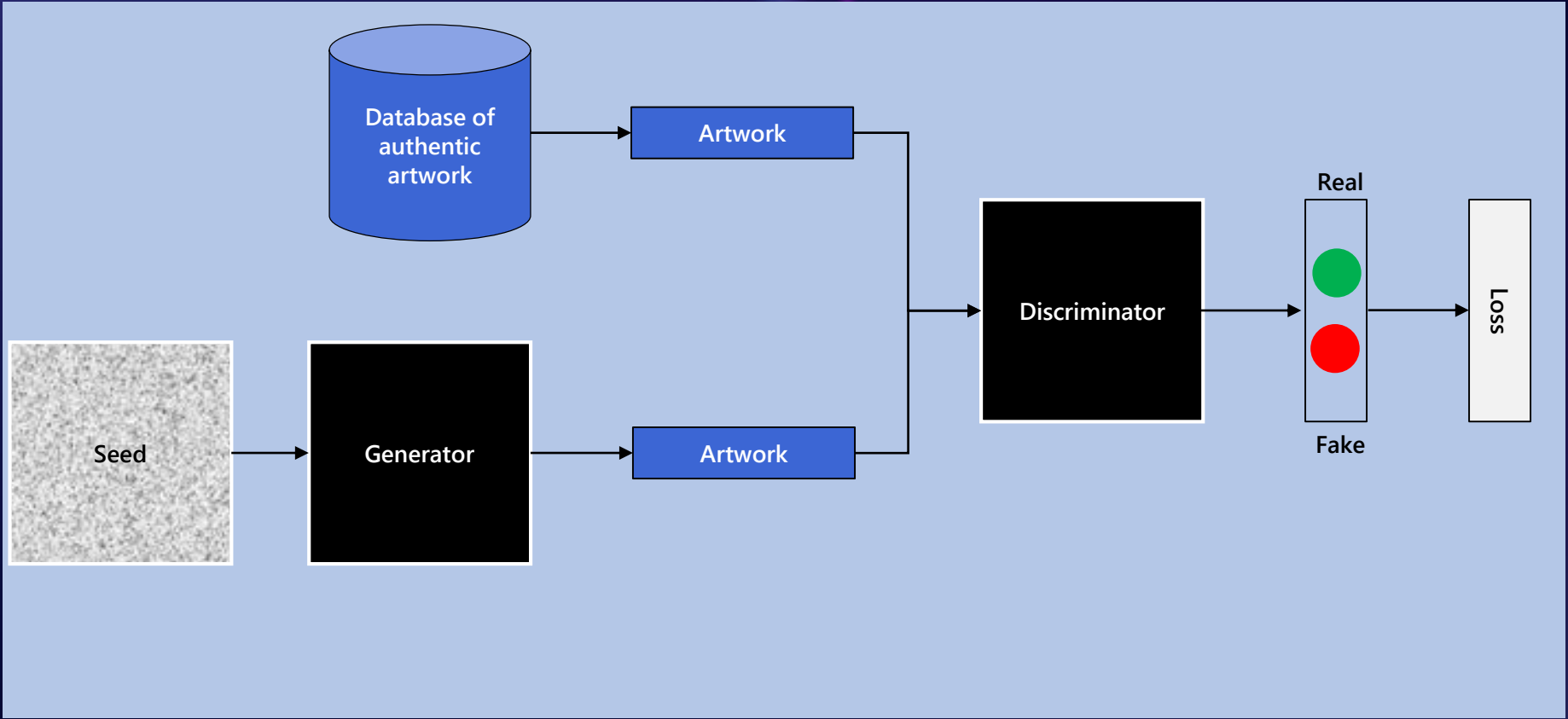
Generative Adversarial Networks

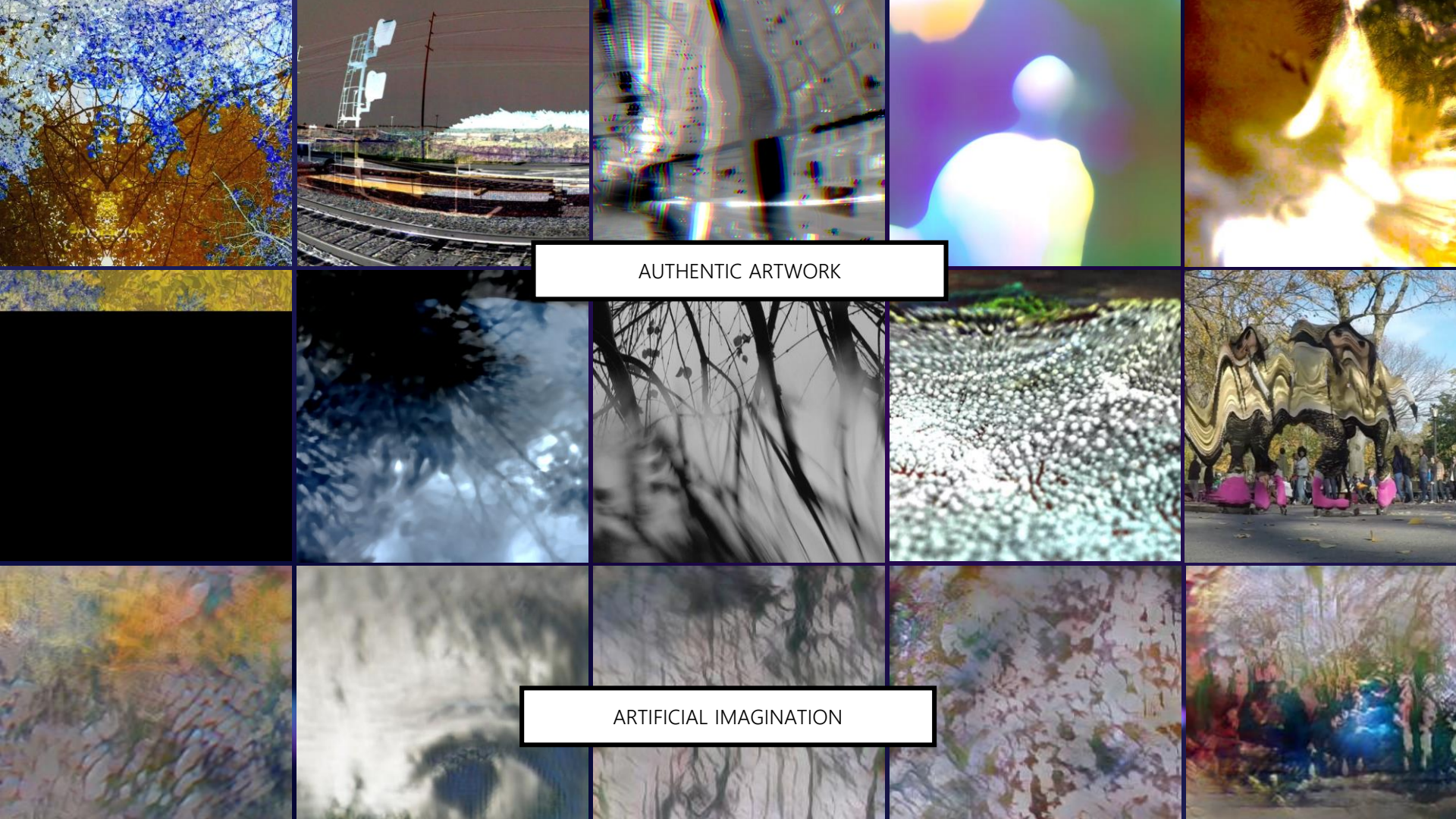
GANs are a branch of deep learning composed of 2 separate neural networks competing each other: the generator and the discriminator, in a zero-sum game framework.

Their goal is to generate data points that are magically similar to some of the data points in the training set.

[Goodfellow, et al., 2014]


Training set of images
Sample
Seed
Generator
Sample
Discriminator
Real
Fake
Loss


Database of authentic artwork
Artwork
Seed
Generator
Artwork
Discriminator
Real
Fake
Loss


AUTHENTIC ARTWORK
ARTIFICIAL IMAGINATION

ARTIFICIAL IMAGINATION

理解语言，用绘画来表达

一只红羽毛白肚子的短咀小鸟

2018

【Xu, Zhang, Huang, Zhang, Gan, Huang, He, CVPR2018】

Stacked Attention for GANs

- Propose Attention GANs to improve image generation

- Goals
  - Improve the quality of generated images
  - Improve the interpretability of GANs
  - Stabilize the training of GANs

- Solution: Attention Generative Adversarial Networks (AttnGANs)
  - Propose the generative networks with stacked attention to generate images from low-to-high resolutions at different stages.
  - Propose a deep attention multimodal similarity model to learn visually-discriminative word features in an semi-supervised manner.

[Xu, Zhang, Huang, Zhang, Gan, Huang, He, ''AttnGAN: Fine-grained text to image generation with Attentional GANs'', CVPR, 2018]

AttnGAN: 智能绘画机器人

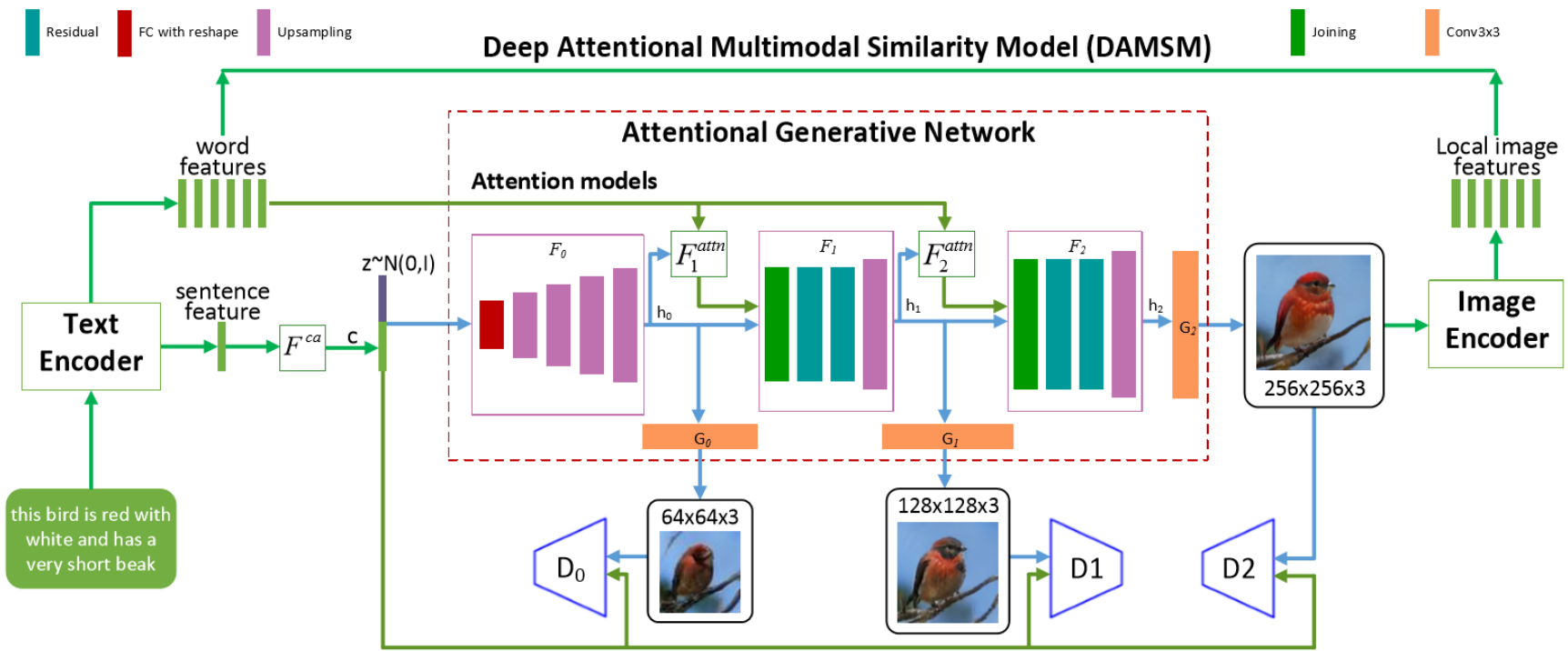
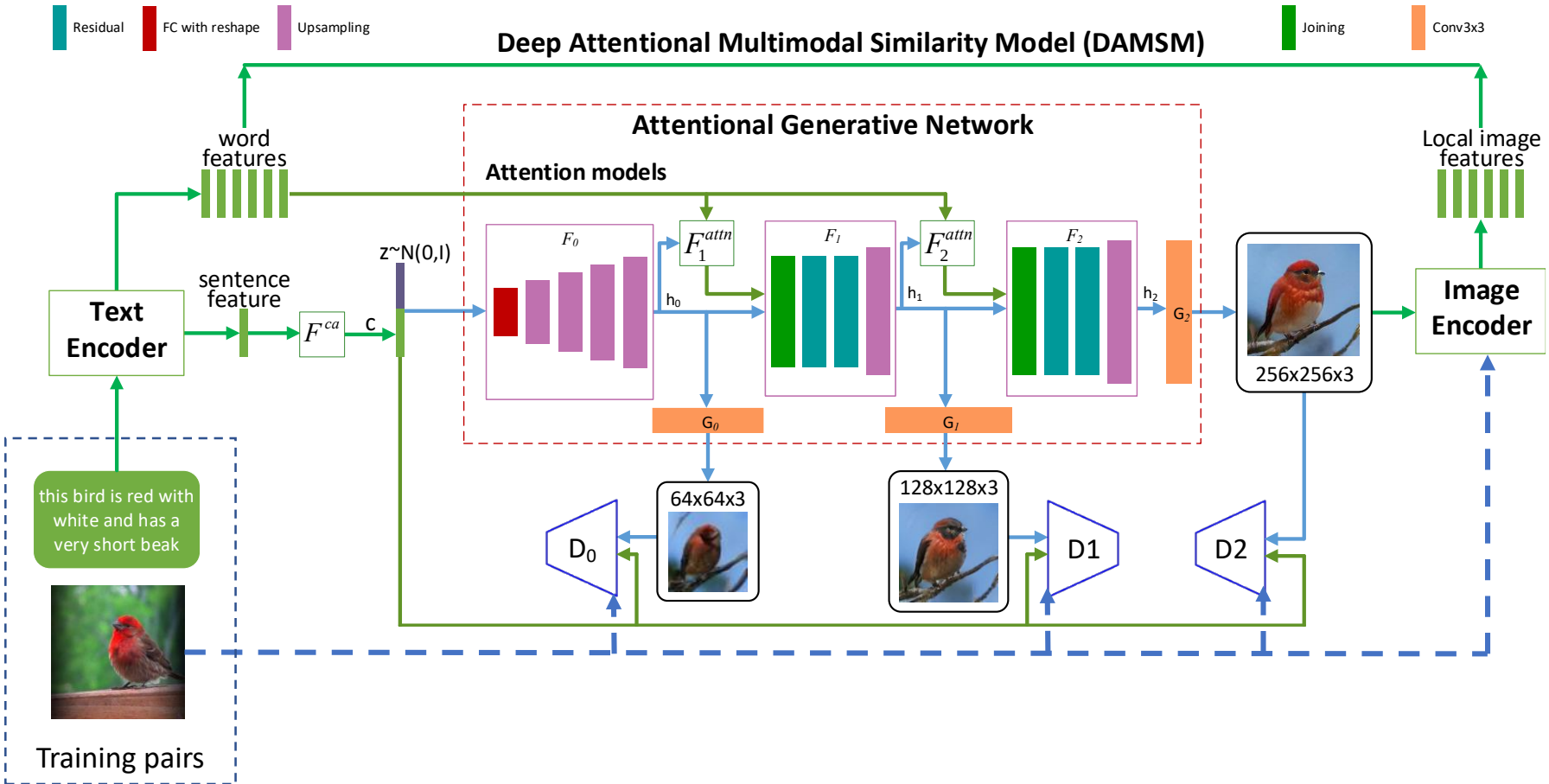


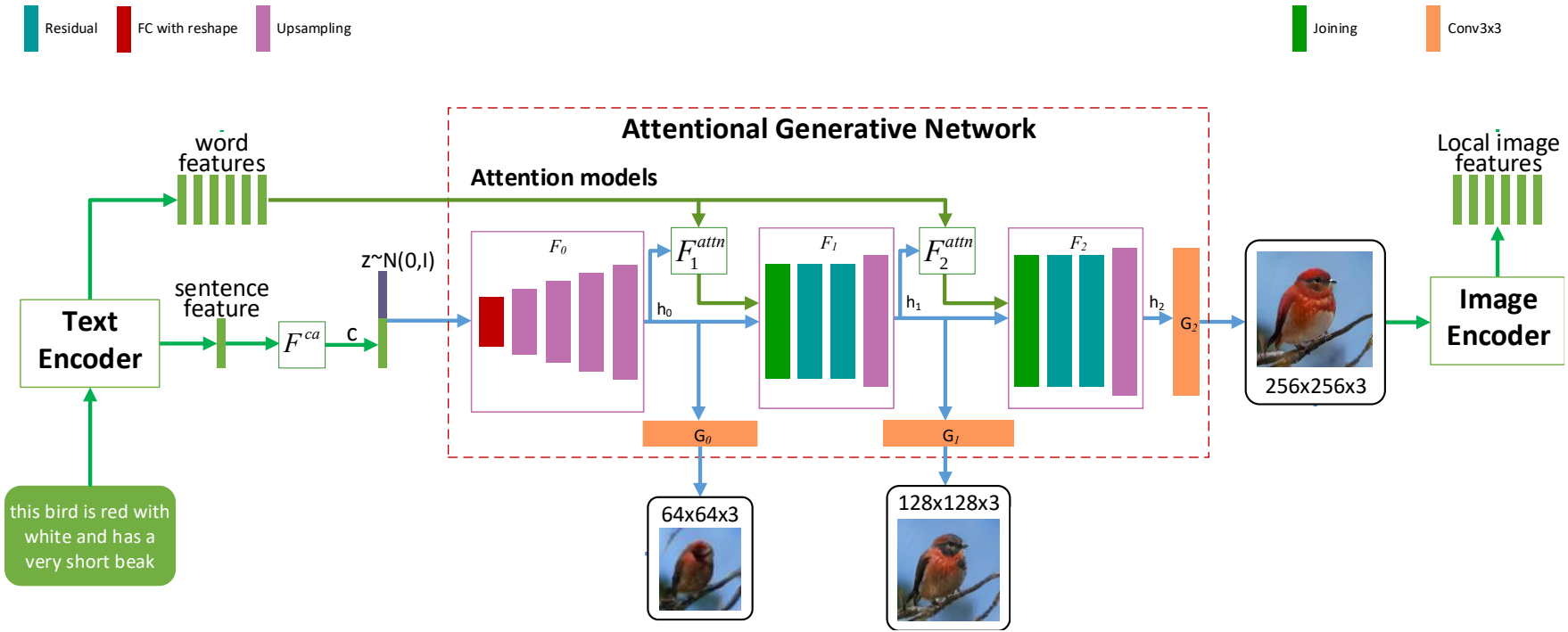
Figure 2. The architecture of the proposed AttnGAN. Each attention model automatically retrieves the conditions (*i.e.*, the most relevant word vectors) for generating different sub-regions of the image; the DAMSM provides the fine-grained image-text matching loss for the generative network.

【Xu, Zhang, Huang, Zhang, Gan, Huang, He, CVPR2018】

AttnGAN: Overview



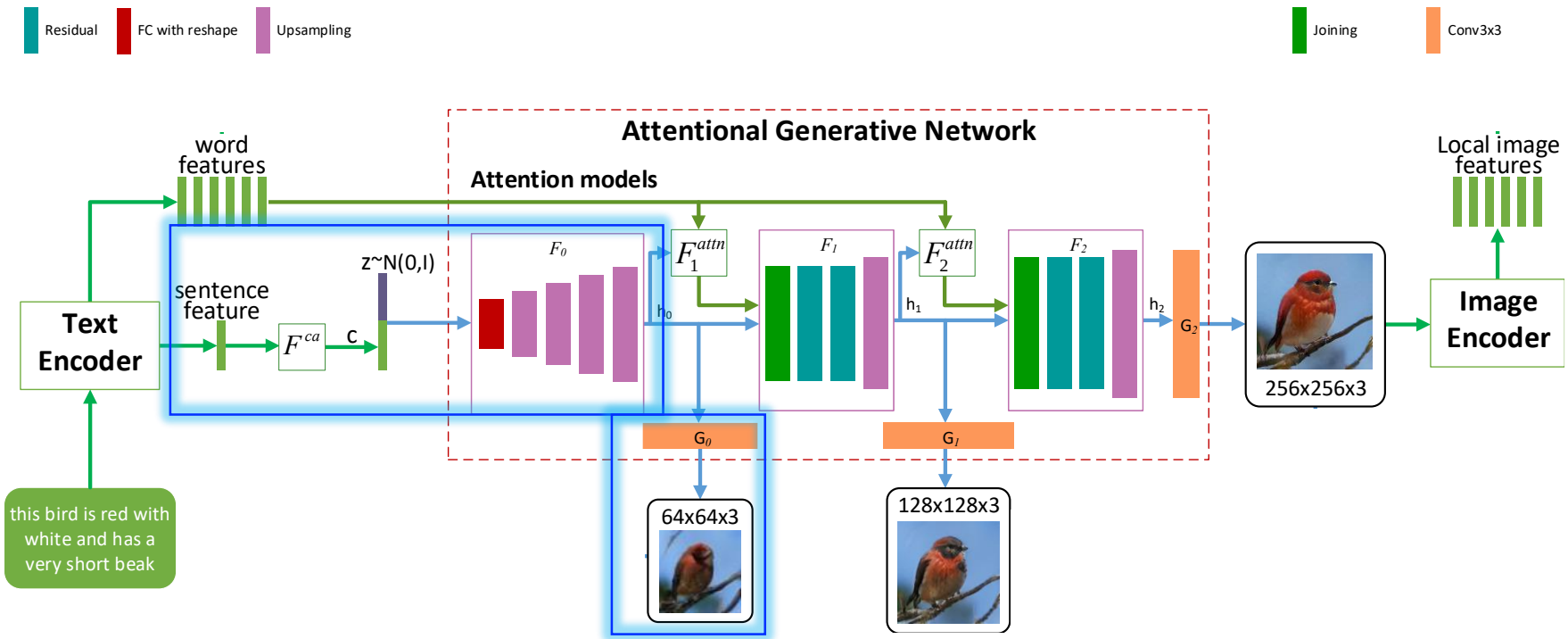
AttnGAN: Attentional Generative Network



Attentional Generative Network:

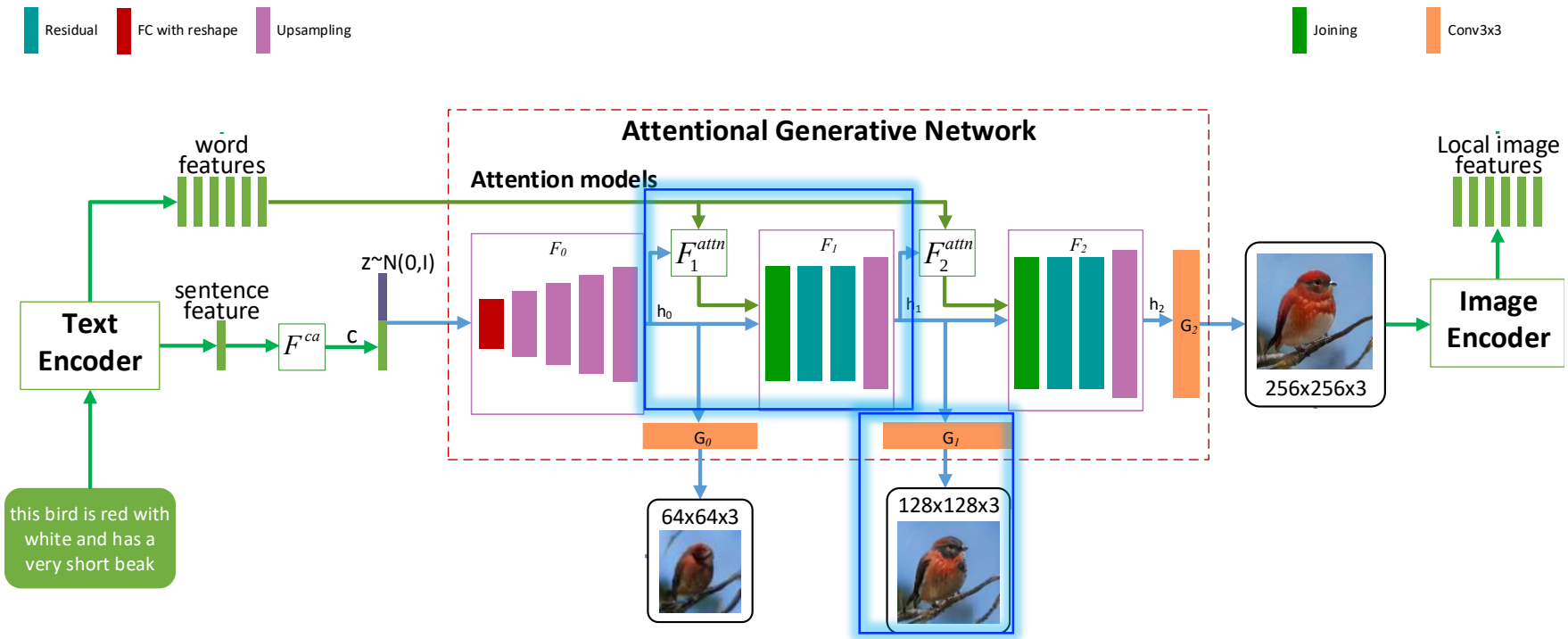
- Takes multi-level conditions (global-level sentence feature and fine-grained word features) as input.
- Generates multi-scale images stage by stage by focusing on different words.

AttnGAN: Attentional Generative network



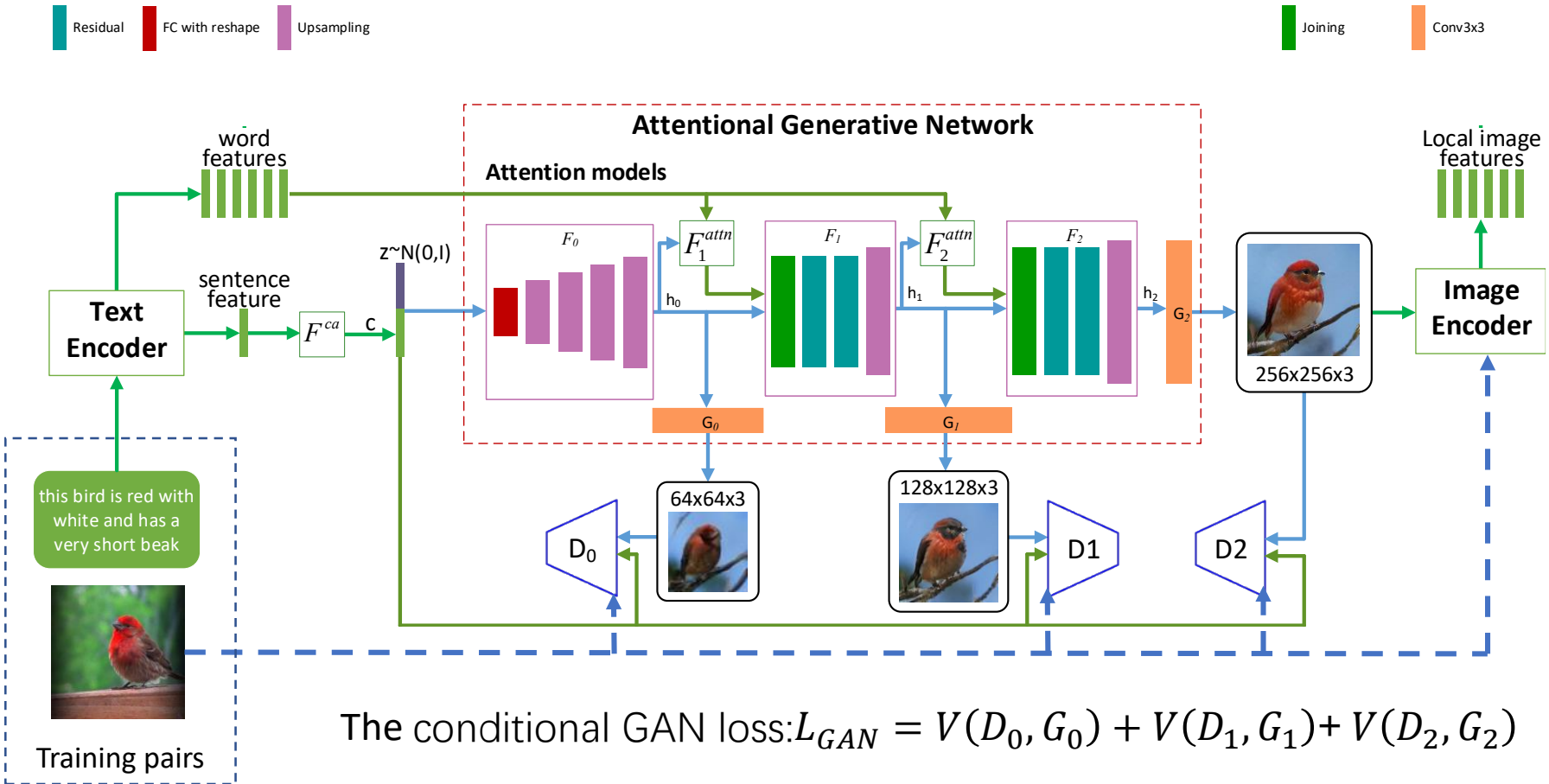
- In the first stage:
  - based on the sentence feature, the image with basic color and shape is generated by generator G_0;
  - hidden features h_0 are decoded from the sentence feature.

AttnGAN: Attentional Generative network



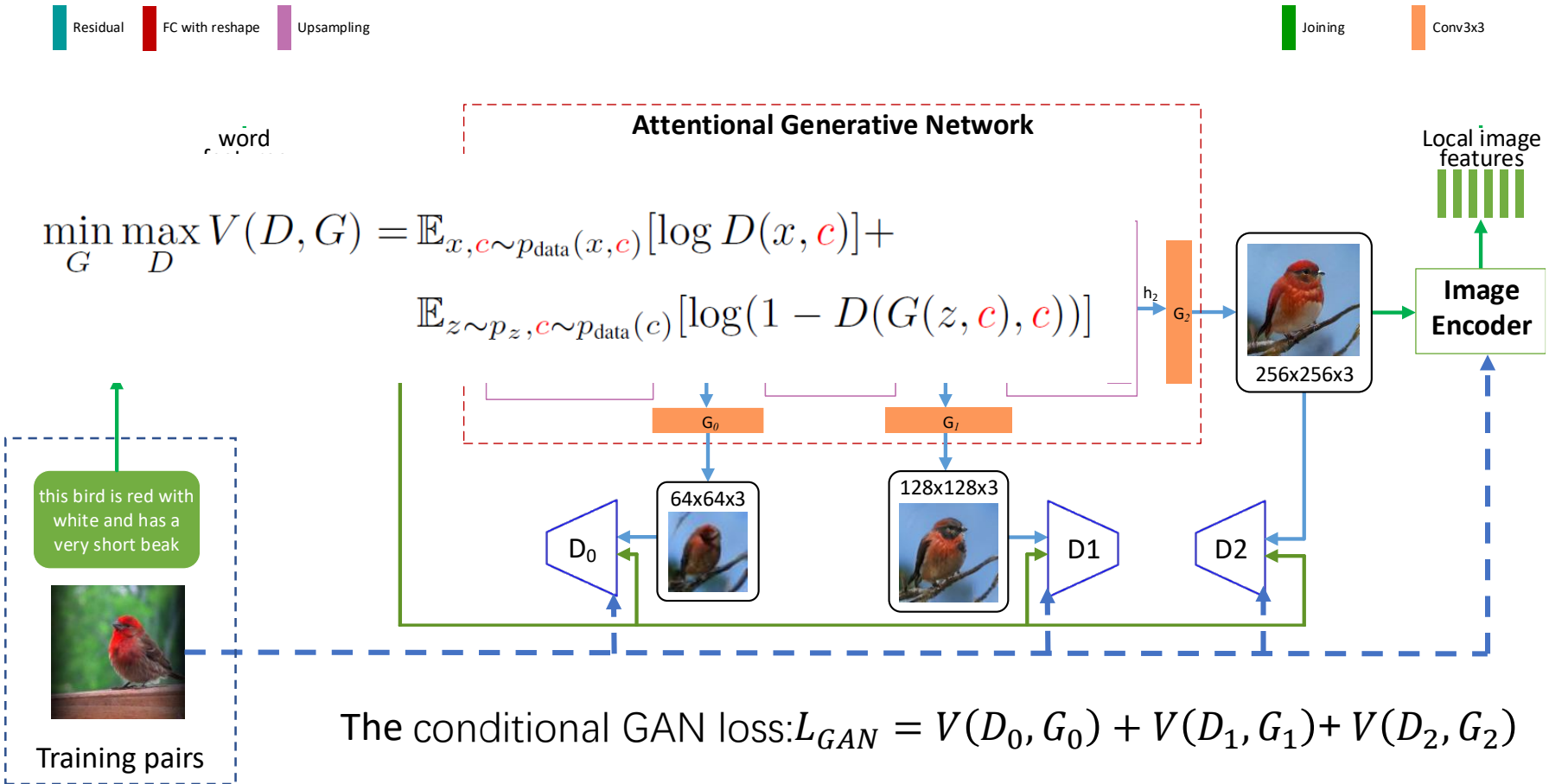
- In following stages, attention models are built.
  - For each region feature of previous generated image, compute its word-context vector.
  - Concatenate previous image region features (e.g., h_0) and word-context vectors to generate the new image.

AttnGAN: the conditional GAN loss



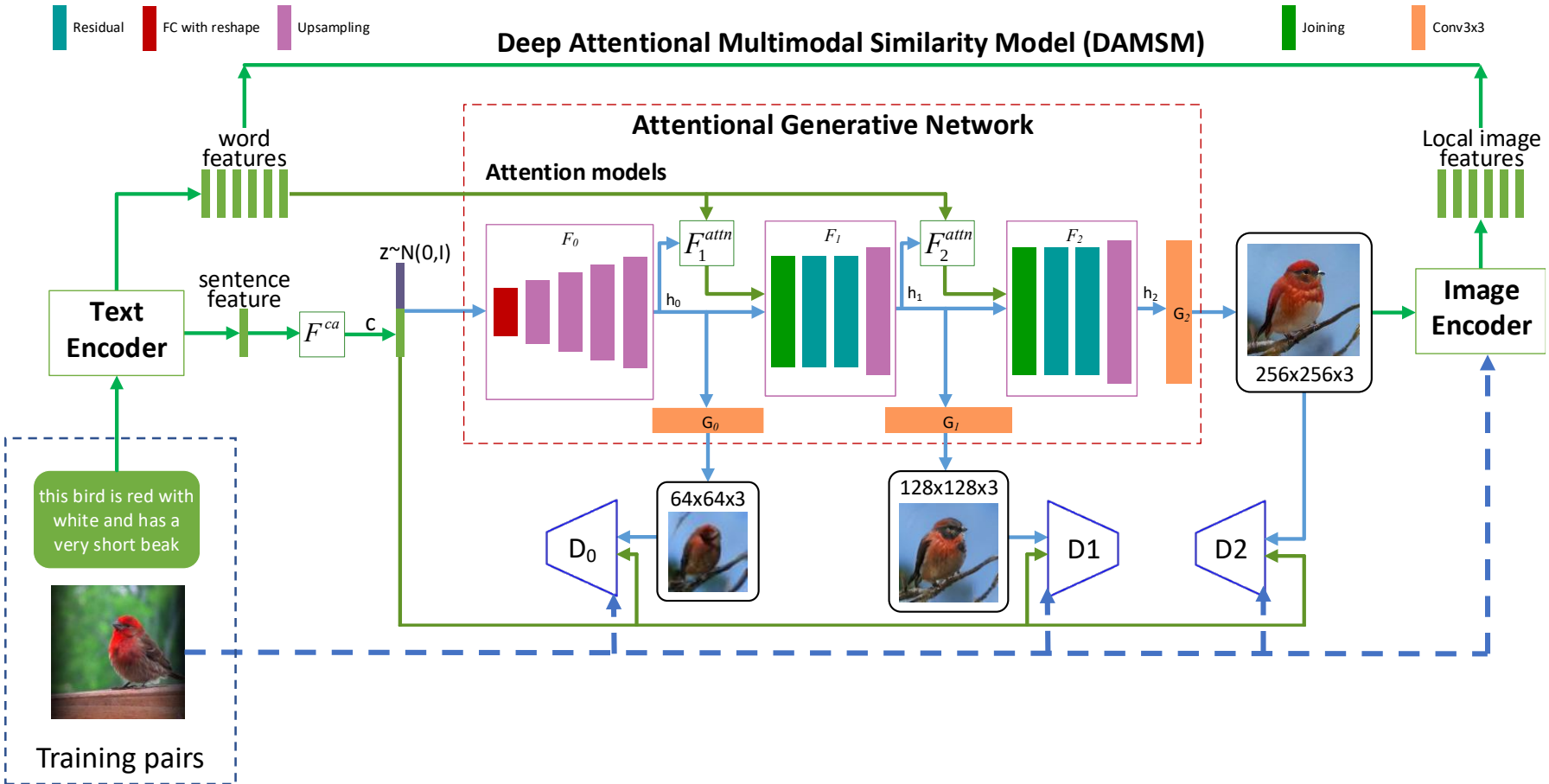
The conditional GAN loss: $L_{GAN} = V(D_0, G_0) + V(D_1, G_1) + V(D_2, G_2)$

AttnGAN: the conditional GAN loss



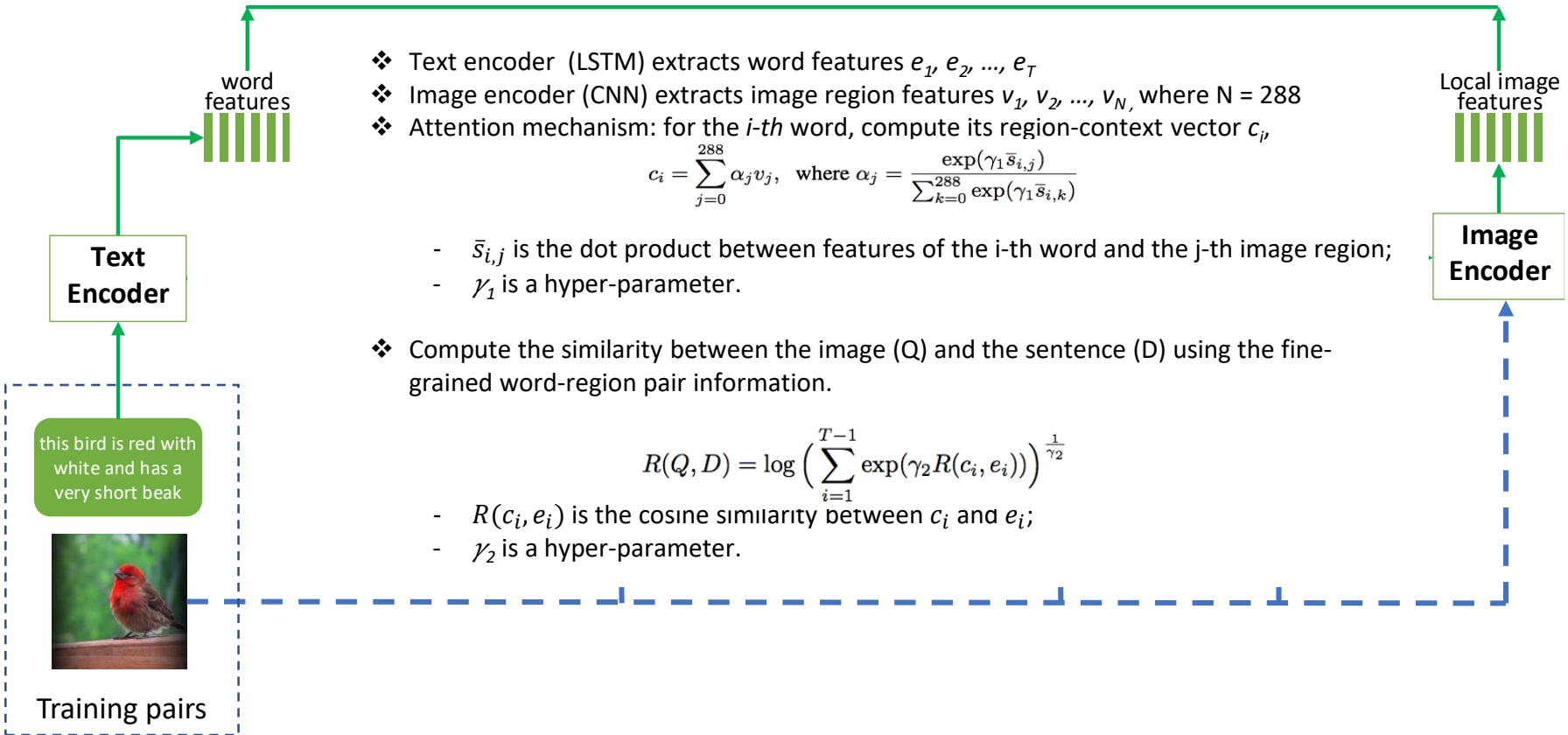
The conditional GAN loss: $L_{GAN} = V(D_0, G_0) + V(D_1, G_1) + V(D_2, G_2)$

AttnGAN: Overview



AttnGAN: DAMSM sub-network

Deep Attentional Multimodal Similarity Model (DAMSM)



- Text encoder (LSTM) extracts word features $e_1, e_2, \ldots, e_T$
- Image encoder (CNN) extracts image region features $v_1, v_2, \ldots, v_N$, where N = 288
- Attention mechanism: for the *i-th* word, compute its region-context vector c_i,

$$c_i = \sum_{j=0}^{288} \alpha_j v_j, \quad \text{where } \alpha_j = \frac{\exp(\gamma_1 \bar{s}_{i,j})}{\sum_{k=0}^{288} \exp(\gamma_1 \bar{s}_{i,k})}$$

  - $\bar{s}_{i,j}$ is the dot product between features of the i-th word and the j-th image region;
  - γ_1 is a hyper-parameter.

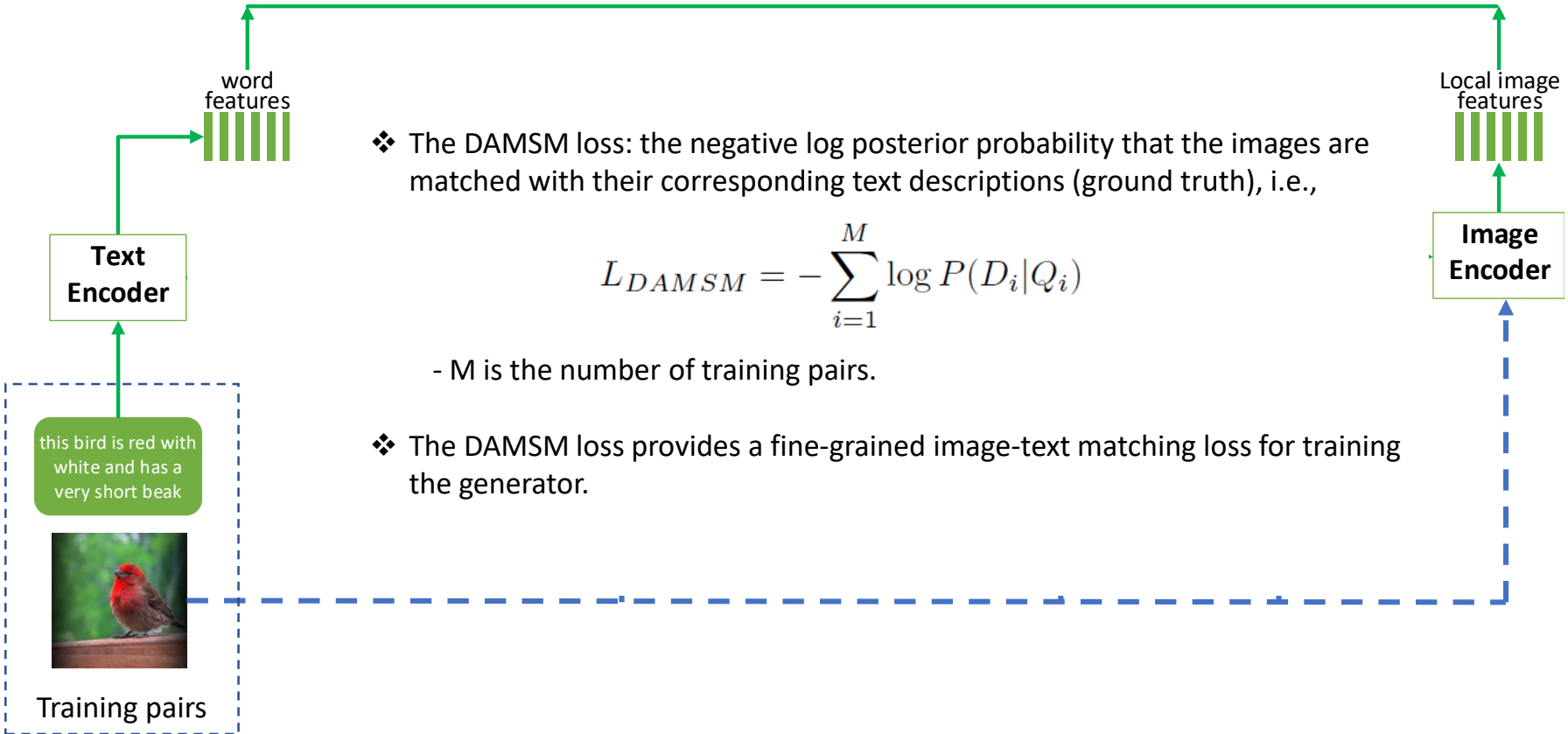
- Compute the similarity between the image (Q) and the sentence (D) using the fine-grained word-region pair information.

$$R(Q, D) = \log\left(\sum_{i=1}^{T-1} \exp(\gamma_2 R(c_i, e_i))\right)^{\frac{1}{\gamma_2}}$$

  - $R(c_i, e_i)$ is the cosine similarity between c_i and e_i;
  - γ_2 is a hyper-parameter.

AttnGAN: DAMSM sub-network

Deep Attentional Multimodal Similarity Model (DAMSM)



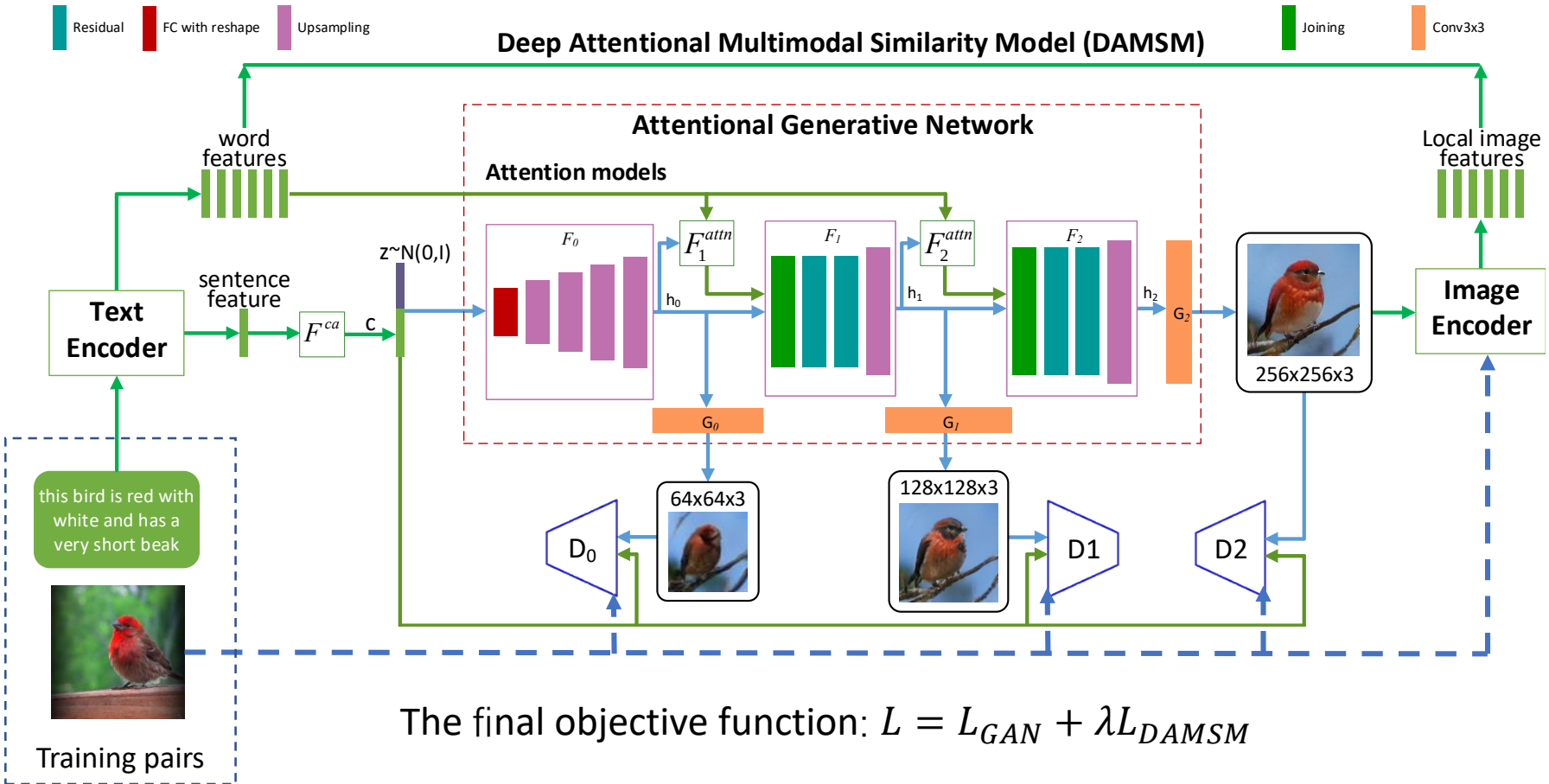
- The DAMSM loss: the negative log posterior probability that the images are matched with their corresponding text descriptions (ground truth), i.e.,

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

  - M is the number of training pairs.

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

AttnGAN: Overview

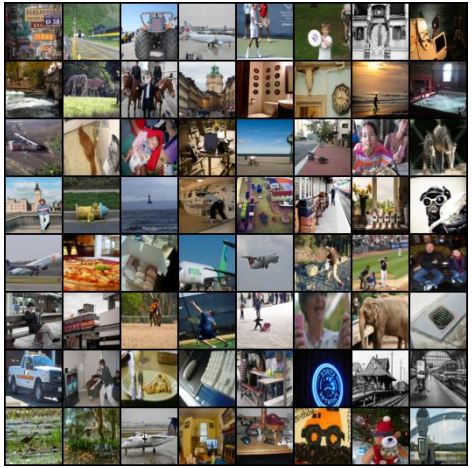


The final objective function: $L = L_{GAN} + \lambda L_{DAMSM}$

Experiments

- Datasets

Datasets	CUB-2011		MS-COCO	
	train	test	train	test
# samples	8,855	2,933	80,000	40,000
caption/ image	10	10	5	5



- Evaluation metrics:
  - **Inception score** reflects the quality and diversity of the generated images.
  - **R-precision** reflects whether the generated images are well conditioned.
  - **Qualitative examination** visualizes the top-5 most attended words.

Comparison with previous methods

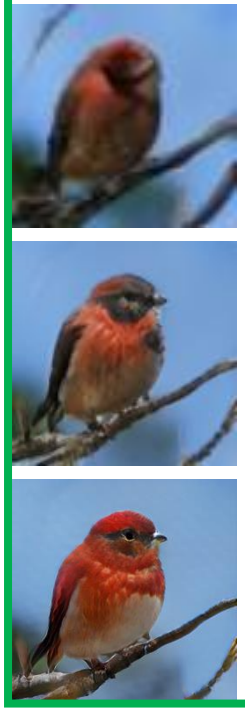
- On CUB dataset, our AttnGAN achieves 4.36 inception score, which significantly outperforms the previous best inception score of 3.82.
- On the COCO dataset, our AttnGAN boosts the best reported inception score from 9.58 to 25.89, a 170.25% improvement relatively.

Dataset	GAN-INT-CLS [1]	GAWWN [2]	StackGAN [3]	StackGAN-v2 [4]	PPGN [5]	Our AttnGAN
CUB	$2.88 \pm .04$	$3.62 \pm .07$	$3.70 \pm .04$	$3.82 \pm .06$		**$4.36 \pm .03$**
COCO	$7.88 \pm .07$	\	$8.45 \pm .03$	\	$9.58 \pm .21$	**$25.89 \pm .47$**

[1] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In ICML, 2016.
[2] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In NIPS, 2016.
[3] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017.
[4] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. arXiv: 1710.10916, 2017.
[5] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug & play generative networks: Conditional iterative generation of images in latent space. In CVPR, 2017.

Qualitative analysis – CUB attention maps

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



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 eyering that is round

a small red and white bird with a small curved beak

AI根据语言描述创作绘画

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

更多例子

an robot that wants to kill all bugs

I can’t find my horse, I miss him so much

my father won't respond to my mamma via memes

a young girl eating a slice of pizza

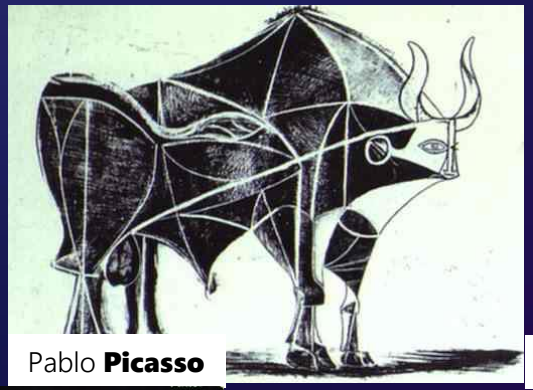
飞速发展的技术领域

[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**]

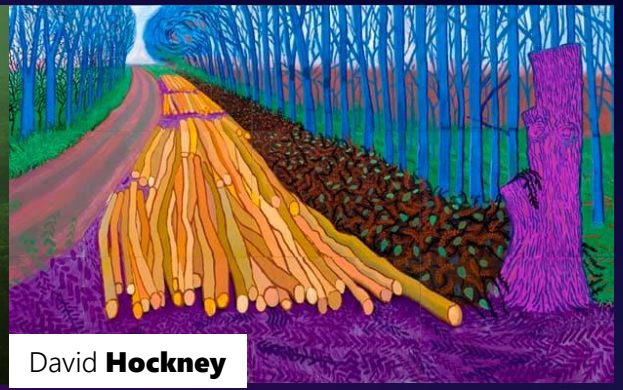
[Brock et al., BigGAN: A New State of the Art in Image Synthesis, ICLR **2019**]

技术改变艺术家的创作方式



Pablo **Picasso**

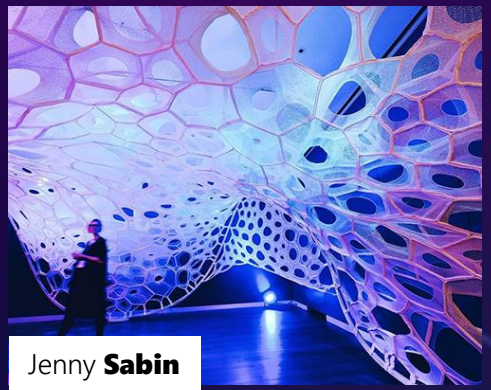
Trevor **Paglen**



David **Hockney**

Andy **Warhol**

James **George**



Jenny **Sabin**

Paint 3D

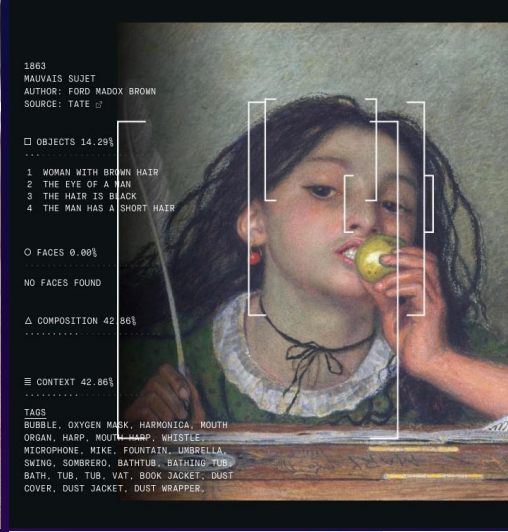
Remix 3D

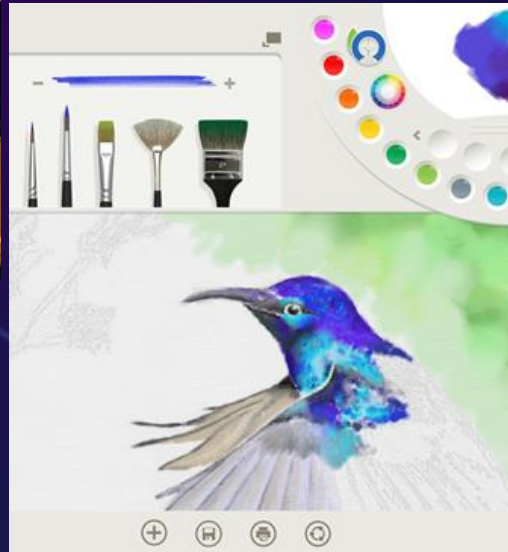
Inking in Office

Creation Tools (Office)

Fresh Paint

MS Arts & Culture





The Google Cultural Institute

Creation Tools (G Docs)

Music Collab (Alex Da Kid)

Movie Trailers (Morgan)

Chef Watson

Cognitive Fashion

Adobe

Adobe Sensei

Project Voco

Scene Stitch

Cloak

Scribbler

AI+Art

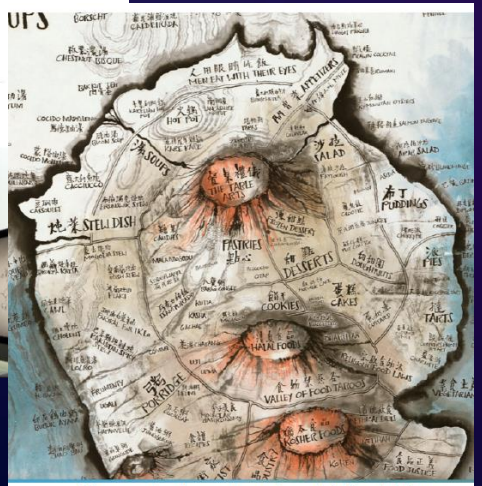
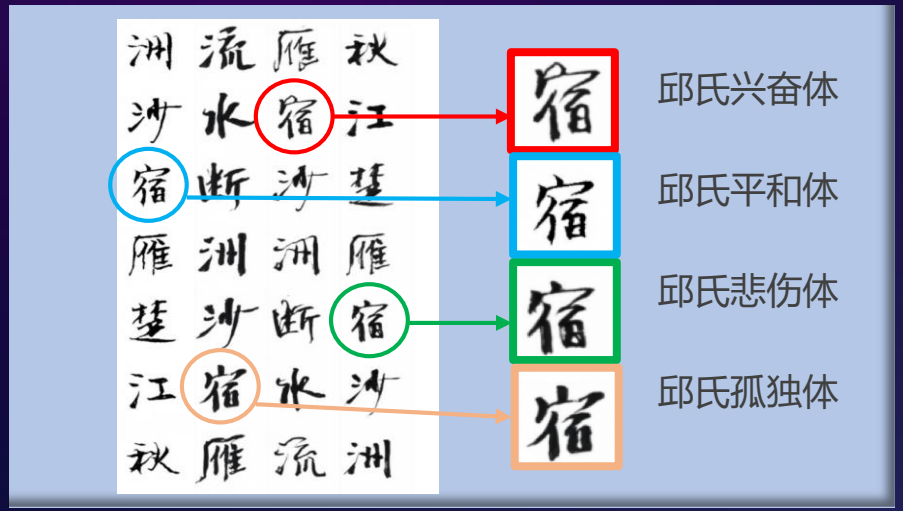


Figure 1: An example of Mind Map



【Liu, et al., 2019】（JDAI与中央美术学院合作）

多模态智能的下一步

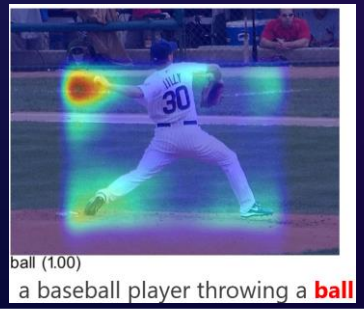
- 融合多个子领域（语言、视觉、语音、知识工程…）
- 跨模态预训练、概念/实体grounding、神经-符号处理机
- 通过复杂跨模态问题驱动基础研究
- 打造更成熟和多元化的实际多模态智能应用

跨越语言和视觉的理解和表达

Universal Chatbot, Digital Assistant, Mixed Reality, …

Multimodal Intelligence: perception, reasoning, and expression across language & vision

Image-to-language



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

人工智能产业化方向探讨

- 智能交互式服务（信息、娱乐、内容、社交、消费）

  传统人力密集型产业，有广阔自动化、智能化空间

  服务即交互：能完成复杂任务的智能人机交互与对话技术，为生活、娱乐、消费、零售、客服、等提供服务

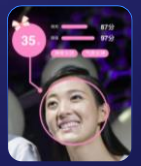
- 智能零售（用户触达、精准营销、C2B/C2M、流程自动化）

  利用视觉、语音、语言等技术提供AI能力助力零售需求

- 智能物联网（音箱、安防、手表、车载）

  人机融合、万物互联、万物智能

人脸/人体分析



- 人脸检测
- 人脸比对与识别
- 特定人物识别

- 人脸防伪检测
- 人脸表情与属性识别
- 人脸搜索
- 人脸生成

- 人体关键点检测
- 人体动作手势识别
- 人体部位解析
- 人体重识别

政要人物识别：NeuHub日均调用量1亿+
CVPR LIP 竞赛人体姿态估计竞赛两项第一 (2018)
人脸检测：国际人脸检测WIDER FACE竞三项第一 (2019)

京灵服务机器人

智能交互屏

蜂巢：城市智能服务空间

智能商业屏

会员识别屏

自助收银机

应用案例

京灵服务机器人

蜂巢：智能服务空间

智能零售 – 门店科技：打造线上线下一体化消费新体验

线下零售科技

智能硬件/IOT

- 会员识别屏
- 智能结算台
- 商业智能屏

开放组件

- 门店小程序(顾客端)
- 京东品牌营销平台(商家端)

线上零售科技

智能分析

财务分析	运营分析	用户分析
收入、成本、利润、市场占有率……	销售，活动，品牌，内容，会员，供应链，服务……	用户画像、用户行为分析、用户偏好分析

精准运营

营销	CRM	广告
GMV	用户忠诚度	大众媒体投放
新品运营	用户拉新	精准广告投放
尾货清仓	用户复购	

应用案例

543 铜锣湾Circle-K店

411 长沙湾Circle-K店

打造AI咨询大脑，全面赋能咨询服务产业

金融

教育

媒体

零售

咨询大脑

制造加工

医疗服务

交通运输

更多

人机交互智能服务体验

呈现：对话式体验

技术：多轮交互与博弈

咨询

交互

关怀

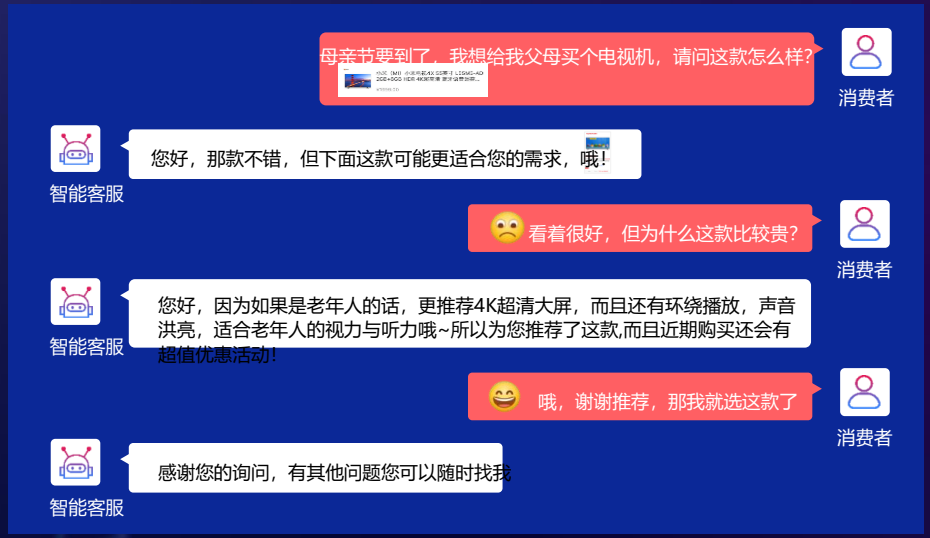
多轮对话

情感智能

认知智能

博弈与决策

对话智能技术：支撑对话式咨询服务



- **多轮对话** - 上下文跟踪；对话管理
- **情感智能** - 情绪识别；同理心反馈
- **认知智能** - 语义理解；知识推理
- **博弈和决策** - 理解用户需求；推荐最佳解决方案全局、长期优化；完成交易

基础支撑技术：

计算机视觉　语音　语义　知识图谱　人机对话　深度学习

京东AI通用对话技术平台

产品

- 在线对话机器人
- 语音应答机器人
- 语音外呼机器人
- 服务质检系统
- 调度大脑
- ...

入口集成

- Web端
- H5端
- 电话系统
- 硬件接入
- 微信接入
- 微博接入

AI能力组件

- 语音识别
- 自然语言生成
- 语音合成
- IR-Bot对话框架
- 自然语言理解
- Task-Bot对话框架
- ...

客户运营后台

- 资源位配置
- 数据可视化
- 知识库管理

开发者平台

- 企业信息接入
- 对话流程搭建
- 领域最佳对话流程克隆

模型、知识和数据沉淀

- 客服ASR模型
- 零售-售前、中、后对话模型
- 零售-商城NER模型
- 政务-政策咨询对话模型
- 医疗-医疗服务分流NLU模型
- 大家电商品图谱
- ...

算法工具集

- 语料清洗工具
- 标注/训练平台
- 数据爬虫

语音识别质检

传统人工质检	语音质检系统
抽检不足2%	全量质检
服务质量不一致	实时反馈 升级处理
风险信息难发现	风险精准识别
热点信息被遗漏	热点智能预测

呼叫中心　政务大厅　银行网点

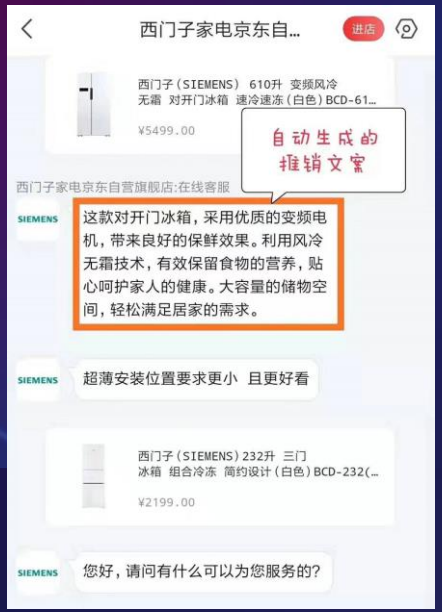
适用场景

智能内容生成

京东AI 为商品写诗，应用于礼品购

AI 写商品推荐短文，人工审核通过率达90%

AI营销客服，自动生成商品亮点文案



咨询大脑 + 零售：零售全链条智能服务

智能客服

智能调度 | 智能导航 | 智能摘要 | 实时辅助应答 | 智能质检 | 智能创事件

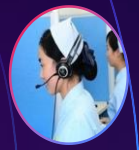
咨询大脑 + 医疗：助力升级“智能+“服务型医院

智能咨询

智能分诊

智能随访

智能客服大脑



智能工作台

智能硬件

咨询大脑 + 媒体

人民日报
PEOPLE'S DAILY
习近平在内蒙古考察，强调了这些大事
全球学生暑期可免票游江西170余家4A级以上景区
一天4趟高铁因旅客吸烟遭降速 7名烟民领罚
JDAI 京东人工智能
TiKr

京东
京东读书
视听新感受

京鱼座AI音箱 C1
连接你的热爱
199元

JDAI 京东人工智能
超芯科技

京东AI：开放能力版图

服务

人脸识别
- ✓ 人脸对比
- ✓ 人脸搜索
- ✓ 人脸检测与属性分析
- ✓ 活体检测
- ✓ 自拍人像抠图

人体识别
- ✓ 人体关键点检测
- ✓ 人体检测

OCR
- ✓ 通用图片识别
- ✓ 身份证识别
- ✓ 行驶证识别
- ✓ 增值税发票识别
- ✓ 银行卡识别

图像理解
- ✓ 拍照购
- ✓ 智能鉴黄
- ✓ 特定人物识别
- ✓ 菜品识别
- ✓ 服装属性标签

语言基础
- ✓ 词法分析
- ✓ 句法分析

语义理解
- ✓ 情感分析
- ✓ 短文本相似度
- ✓ 文本分类
- ✓ 评论观点抽取
- ✓ 敏感词扩充

语音
- ✓ 语音识别
- ✓ 语音合成

产品
- ✓ 服务机器人
- ✓ 蜂巢机器人
- ✓ 人脸属性互动屏
- ✓ 智能摄像机
- ✓ 智能结算台

解决方案
- ✓ 智能对话
- ✓ 智慧门店
- ✓ 智慧入口
- ✓ 电话机器人

NeuHub 京东人工智能开放平台
搜索
中文站
论坛
控制台
登录
注册
前沿动态
产品及服务
应用平台
开发平台
合作伙伴
AI研究院
支持与服务
NeuBot
京灵服务机器人
查看详情
NeuHub
京东人工智能开放平台
城市综合服务机器人
德清地理信息大会
城市综合服务机器人
德清地理信息大会
城市综合服务机器人
德清地理信息大会
城市综合服务机器人
德清地理信息大会

京东入选国家人工智能开放创新平台 领衔智能供应链

市政场景

零售场景

客服场景

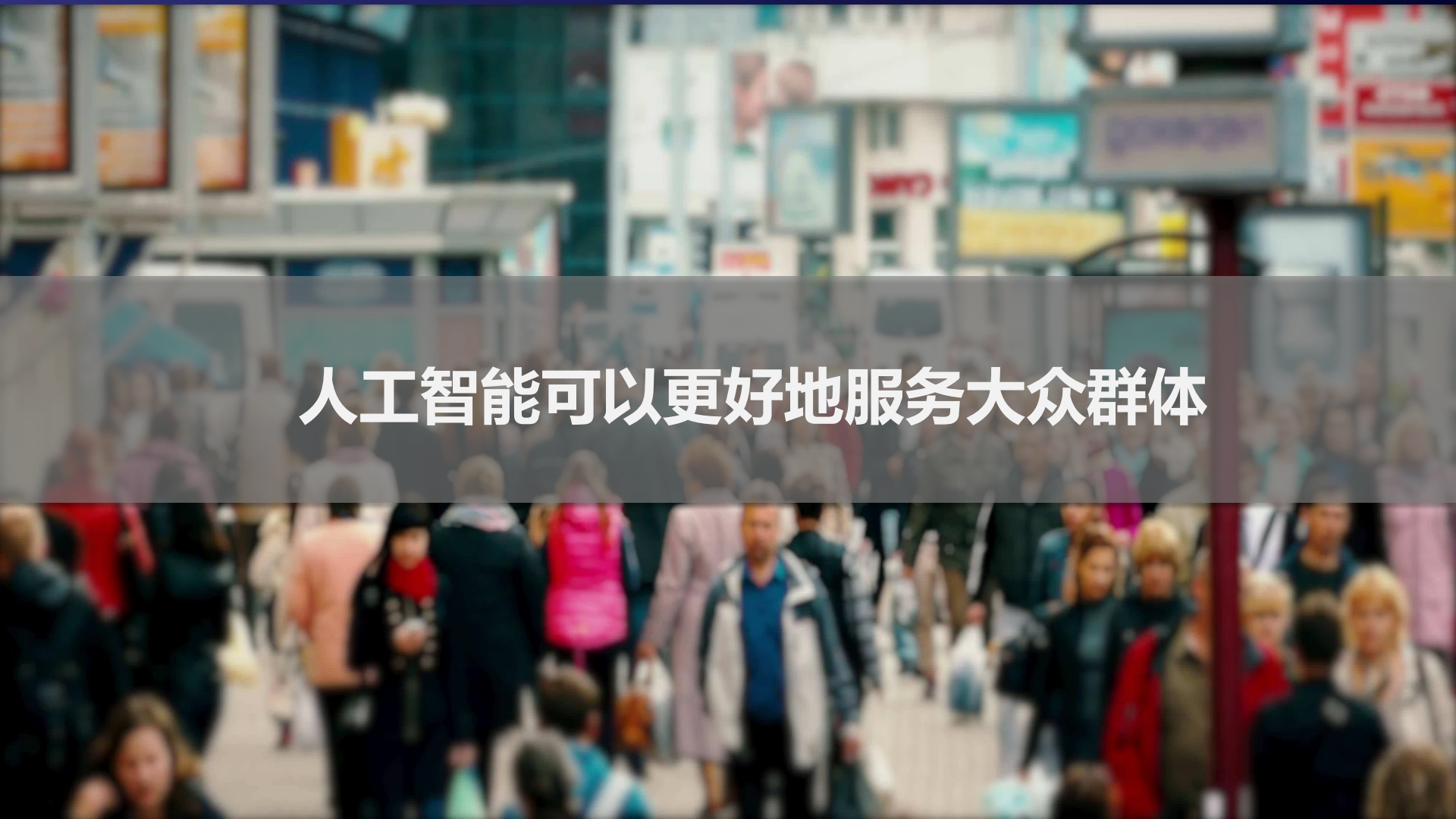
医疗场景

技术研究

产业应用

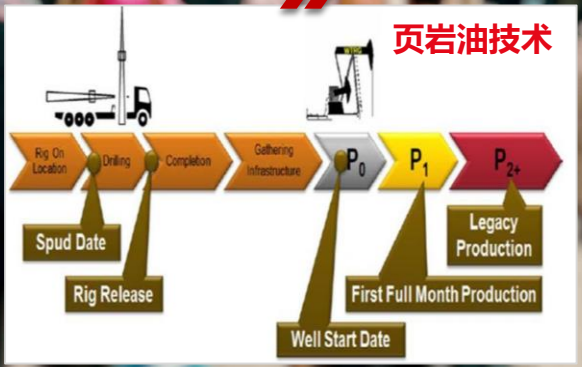
开放生态

NeuHub，京东AI打造的共建、共享、共用的人工智能开放平台



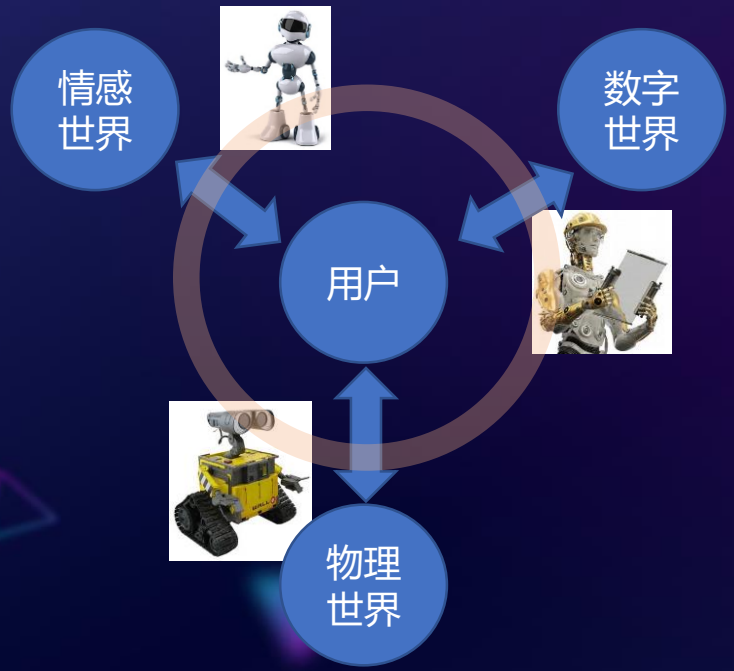
人工智能可以更好地服务大众群体

通过降低成本，AI驱动的服务将挖掘出新的商业机会



人机融合、多模态智能服务的产业时代

服务即对话：多模态、大规模开放领域、具有常识和情感、能完成复杂任务的对话系统是推动下一代人工智能产业的核心技术。



感知智能

认知智能

通用智能

分级	目标
低级智能对话	对用户简单意图进行识别并给出预设答案。
初级智能对话	能识别复杂意图，联系上下文给出回答。
中级智能对话	根据用户问题及情绪完成个性化多轮对话，协助用户达成目标。
高级智能对话	能对多模态信息进行推理，自主判断，并组织语言与用户沟通。具备自我学习能力。
通用智能对话	能基于一切信息开展自我学习，自我适应，及自我创新。在复杂问题领域达到人类水平。

THANKS

Xiaodong.He@JD.COM