厦门大学
媒体分析与计算实验室

福建省可信人工智能
分析与应用工程研究中心

认知驱动的视觉-语言协同分析

汇报人：孙晓帅 博士

厦门大学

2021年6月11日



汇报提纲

XIAMEN UNIVERSITY

一、引子

二、视觉 → 语言

三、语言 → 视觉

四、多特征-多任务协同

五、可控性

六、总结

01 引子
Introduction



研究背景

XIAMEN UNIVERSITY

科学意义

- 复杂语言功能的形成可能是人类其它感官高度进化的结果（因果关系）
- 图像、视频还有语言文字，都是我们希望计算机能够理解的对象
- 将图文联合起来进行研究，符合客观的认知规律

应用需求

- 文本、图像、视频是互联网上存量最大的数据，需研发相关算法有效管理它们

网络安全需求

↓

跨媒体内容分析与管控









研究背景

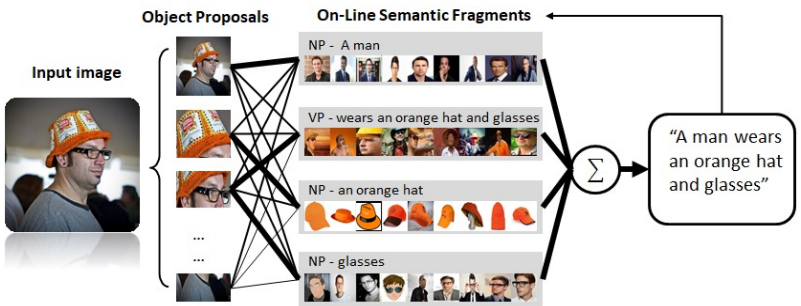
XIAMEN UNIVERSITY

CV领域的视觉-语言跨媒体分析研究

- 图文匹配
- 看图说话
- 视觉问答
- 听文作图
- 指向检测

……

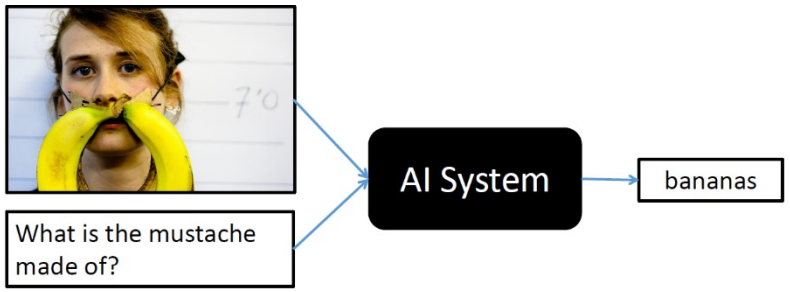
图文匹配
(Image-Text Matching)



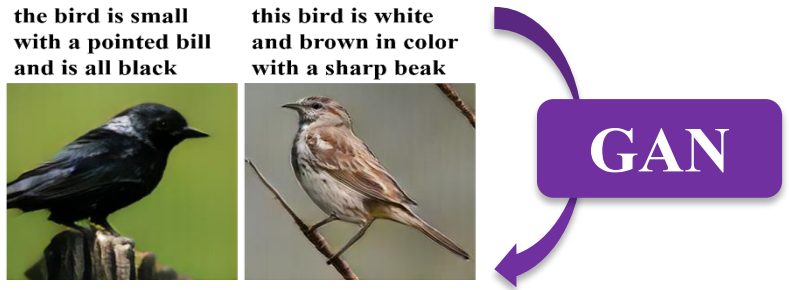
看图说话
(Image/Video Captioning)



视觉问答
(Visual Question Answering)



听文作图
(Language-based Image Generation)



研究背景

XIAMEN UNIVERSITY

课题组的一些相关工作积累

- **视觉描述**

  MM'17（结构描述）AAAI'17（开集图文匹配）, CVPR'18（群组描述）

  NeurIPS'20（多样化描述）, AAAI'21（全局-局部融合，区域-网格协同）

- **视觉问答**

  AAAI'19（知识惯性）, MM'20（网络结构搜索）, TPAMI'20（细粒度训练）

- **注意机制**

  CVPR'12（投影追踪）, AAAI'17（多特征融合）, MM'20（语言引导）

- **指向检测**

  CVPR'20（检测-分割协同）, MM'20（注意串联）

- **对抗攻防**

  ICCV'19（扰动攻击）, CVPR'20（统计黑盒攻击）, MM'20（跨模态攻防）

MAC 代码仓库

02 视觉 → 语言
Vision to Language






Xuying Zhang Yunpeng Luo

RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words. IEEE CVPR 2021

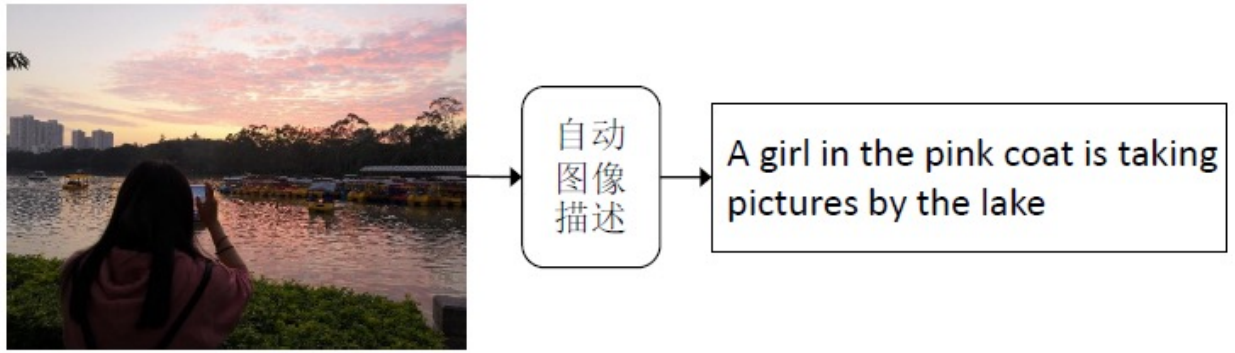
语言的视觉性

语言的视觉性 – 概念

XIAMEN UNIVERSITY

图像描述 (Captioning) 任务

目标：自动生成给定图像内容的自然语言描述

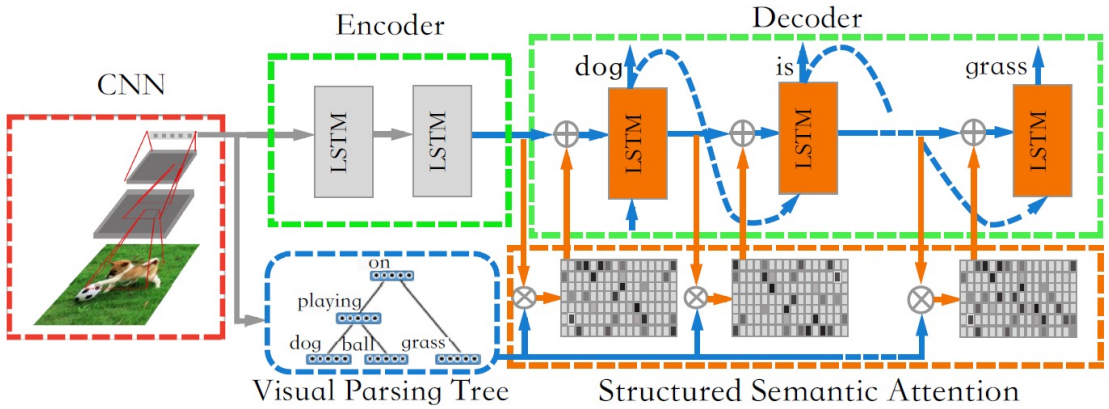


- 连接视觉与自然语言
- 计算机视觉和自然语言处理交叉研究
- 最受关注且最具潜力的人工智能研究任务之一

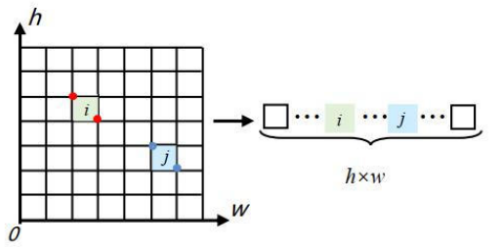
常规自动图像描述	差异化图像描述	风格化图像描述	图像稠密描述	跨域图像描述
指定表达生成	图像段落描述和故事生成	新物体描述生成	基于图像的诗歌创作	图像语音描述
弱监督和无监督图像描述	特定语种的图像描述	视觉问题生成	可控图像描述	带情感的描述生成

语言的视觉性 – 概念

- 图像描述 (Captioning) 任务
  - 目标：自动生成给定图像内容的自然语言描述
  - SOTA：编码器-解码器框架（Transformer及衍生模型）
  - 挑战：1）视觉特征空间关系信息的损失,视觉特征送入解码器时经过展平处理造成空间信息损失
    2）单词同方式处理，没有区分视觉和非视觉词



编码器-解码器框架



空间信息损失



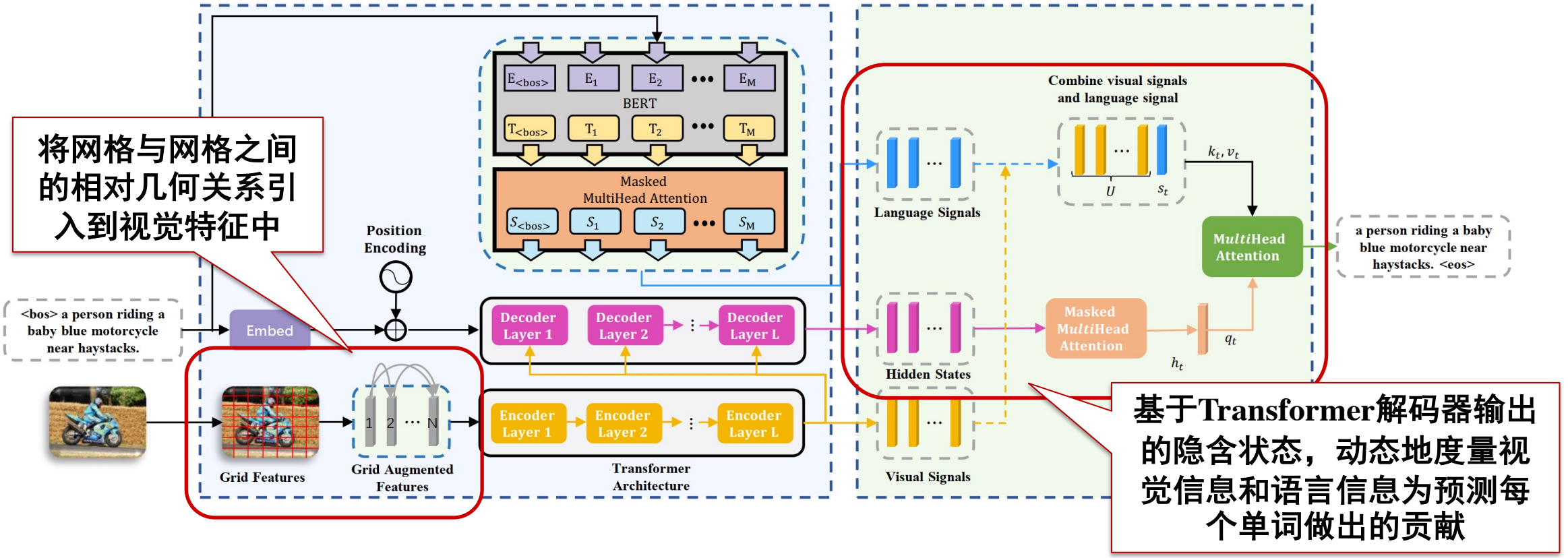
未区分视觉与非视觉词

语言的视觉性 – 方法

XIAMEN UNIVERSITY

模型构建

- 模型：基于网格增强和自适应注意力的图像描述模型
- 特性：解决视觉特征空间关系信息损失和无法区分不同词性单词的问题

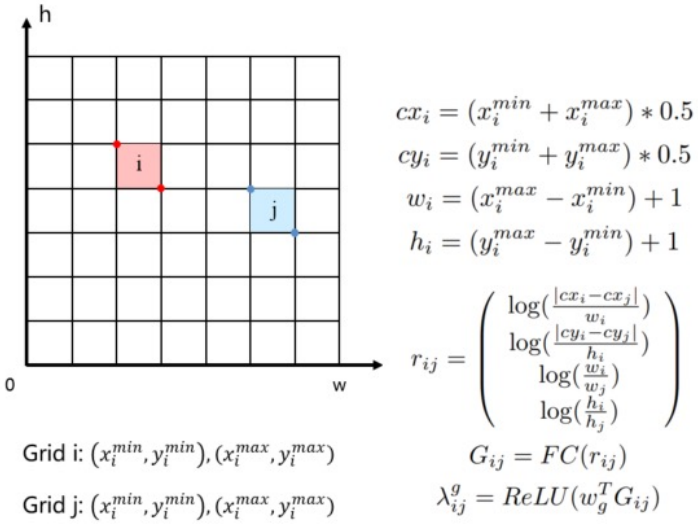


语言的视觉性 – 方法

XIAMEN UNIVERSITY

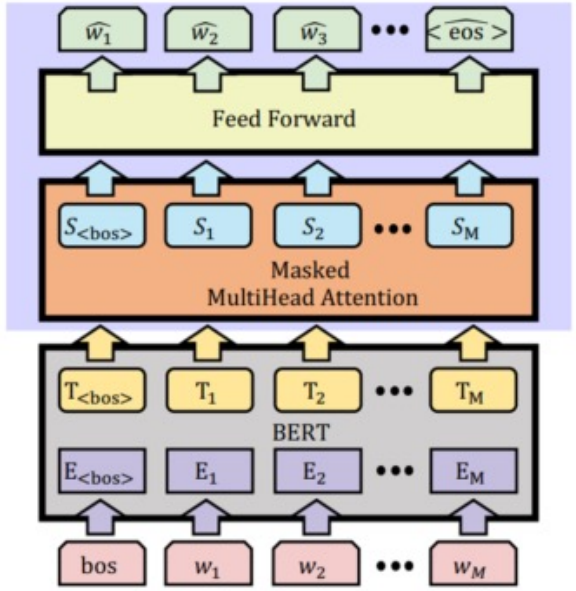
模块解析

- 网格增强(GA)模块
- 基于BERT的预训练语言模型
- 自适应注意力(AA)模块

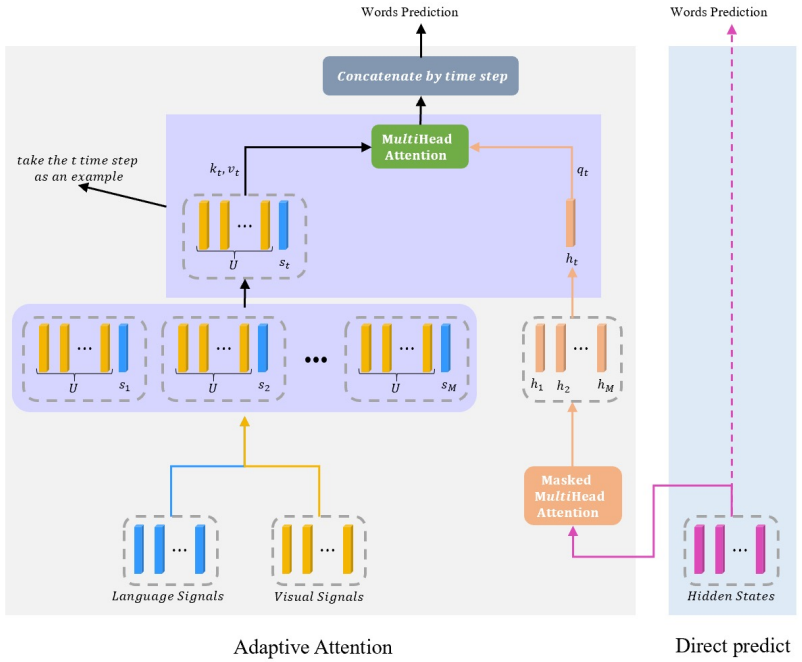


$$cx_i = (x_i^{min} + x_i^{max}) * 0.5$$
$$cy_i = (y_i^{min} + y_i^{max}) * 0.5$$
$$w_i = (x_i^{max} - x_i^{min}) + 1$$
$$h_i = (y_i^{max} - y_i^{min}) + 1$$
$$r_{ij} = \begin{pmatrix} \log(\frac{|cx_i - cx_j|}{w_i}) \\ \log(\frac{|cy_i - cy_j|}{h_i}) \\ \log(\frac{w_i}{w_j}) \\ \log(\frac{h_i}{h_j}) \end{pmatrix}$$
$$G_{ij} = FC(r_{ij})$$
$$\lambda_{ij}^g = ReLU(w_g^T G_{ij})$$

网格增强模块



BERT语言模型



自适应注意模块

语言的视觉性 – 实验

XIAMEN UNIVERSITY

实验结果

Table 3. Performance comparision with the state of the art on the COCO "Karpathy" test split.

Model \ Metrics	B@1	B@4	M	R	C	S
SCST [31]	-	34.2	26.7	55.7	114.0	-
Up-Down [2]	79.8	36.3	27.7	56.9	120.1	21.4
RFNet [16]	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM [43]	80.5	38.2	28.5	58.3	127.6	22.0
SGAE [42]	80.8	38.4	28.4	58.6	127.8	22.1
ORT [12]	80.5	38.6	28.7	58.4	128.3	22.6
AoANet [13]	80.2	38.9	29.2	58.8	129.8	22.4
M^2 Transformer [6]	80.8	39.1	29.2	58.6	131.2	22.6
X-Transformer [28]	80.9	39.7	29.5	59.1	132.8	**23.4**
RSTNet(ResNext101)	**81.1**	39.3	29.4	58.8	**133.3**	23.0
RSTNet(ResNext152)	**81.8**	**40.1**	**29.8**	**59.5**	**135.6**	23.3

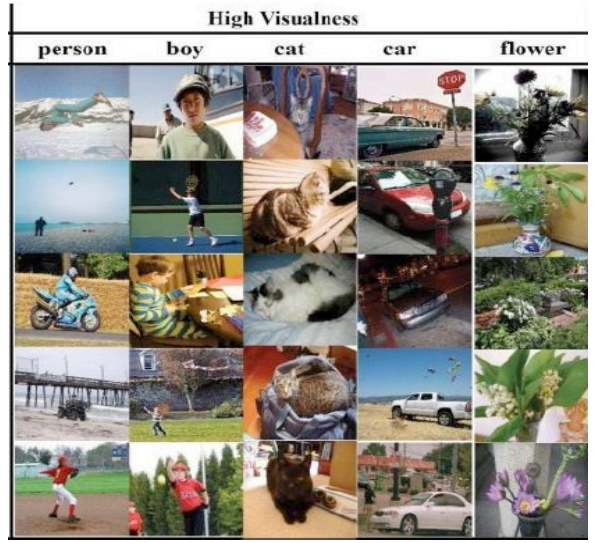
RSTNet在图像描述任务的基准数据集MS-COCO的线上和线下测试中均达到了SOTA

Table 4. Comparing with SOTAs on ResNext101 grid features

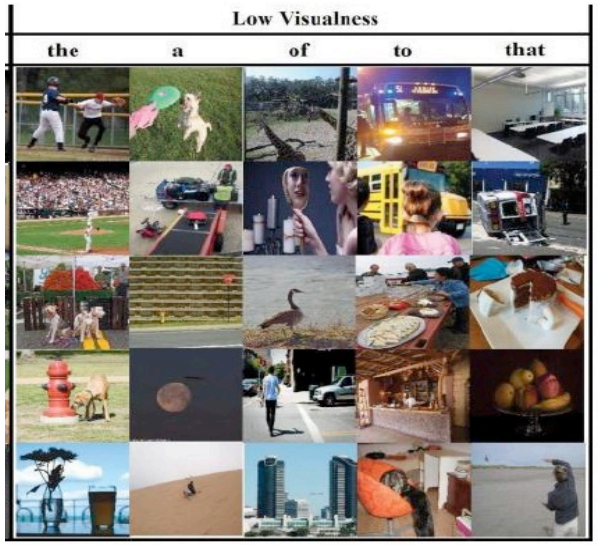
Model \ Metrics	B@1	B@4	M	R	C	S
UP-Down [2]	75.0	37.3	28.1	57.9	123.8	21.6
Transformer [6]	80.9	38.9	29.0	58.5	131.2	22.7
AoA Transformer [13]	80.8	39.1	29.1	59.1	130.3	22.7
M^2 Transformer [6]	80.8	38.9	29.1	58.5	131.8	22.7
X-Transformer [28]	81.0	**39.7**	29.1	**59.0**	130.2	22.8
RSTNet(Ours)	**81.1**	39.3	**29.4**	58.8	**133.3**	**23.0**

Table 5. Leaderboard of the published state-of-the-art image captioning models on the COCO online testing server, where B@N, M, R and C are short for BLEU@N, METEOR, ROUGE-L and CIDEr scores. All values are reported as percentage.

Model \ Metrics	B@1		B@2		B@3		B@4		METEOR		ROUGE-L		CIDEr-D	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST [31]	78.1	93.7	61.9	86.0	47.9	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down [2]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [16]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	37.1	122.9	125.1
GCN-LSTM [43]	80.8	95.9	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE [42]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
ETA [20]	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
AoANet [13]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
M^2 Transformer [6]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Transformer(ResNet-101) [28]	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
X-Transformer(SENet-154) [28]	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
RSTNet(ResNext101)	81.7	**96.2**	66.5	**90.9**	51.8	**82.7**	39.7	72.5	29.3	38.7	59.2	74.2	**130.1**	**132.4**
RSTNet(ResNext152)	**82.1**	**96.4**	**67.0**	**91.3**	52.2	**83.0**	40.0	**73.1**	**29.6**	39.1	**59.5**	74.6	**131.9**	**134.0**



视觉词



非视觉词

03 语言 → 视觉
Language to Vision





Xuying Zhang

Exploring Language Prior for Mode-Sensitive Visual Attention Modeling. ACM MM, 2020.

多模注意机制

多模注意机制 – 概念

XIAMEN UNIVERSITY

■ 视觉注意建模



Bruce & Tsotsos *et al*. 2006
[120 images, 20 viewers]



Torralba *et al*. 2006
[36 images, 24 subjects]



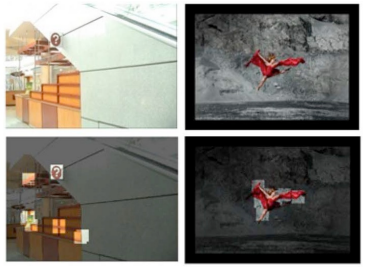
Judd *et al*. 2009
[1K images, 15 viewers]



Subramanian *et al*. 2010
[758 images, 75 subjects]



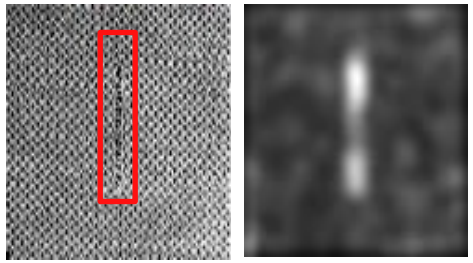
视频编解码



ROI 检测



图像渲染



缺陷瑕疵检测



机器视觉



异常行为检测

多模注意机制 – 概念

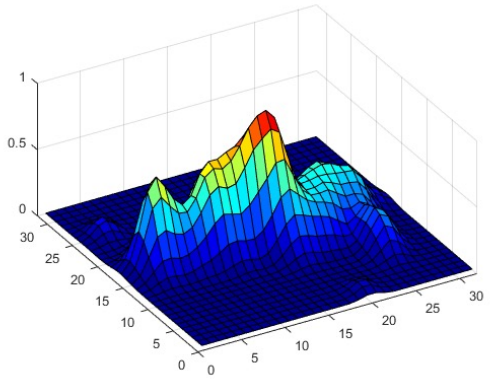
XIAMEN UNIVERSITY

- 语言先验驱动的多模视觉注意
  - 在处理不同风格视觉输入时，注意分布存在偏差，可能采用不同工作模式

Image#1



Fixation Density of #1

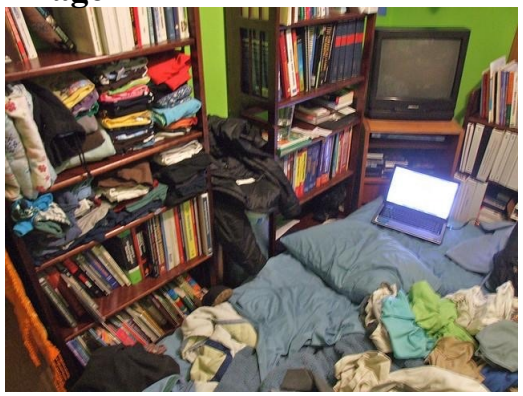


MSCOCO Captions of Image #1

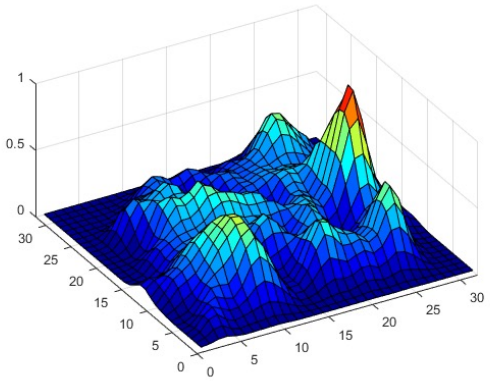
A double decker city bus driving down the street.
A double decker bus waits at a bus stop.
A large double decker bus driving down a city street.
A double decker bus parked on the street.
A double decker bus going down the street.

High Caption Consistency (HCC)

Image#2



Fixation Density of #2



MSCOCO Captions of Image #2:

A laptop is on the bed in a room full of books and clothes.
A messy room with clothes and a computer.
A messy bedroom with the bed on the floor.
A bed covered in lots of clutter with an open laptop computer on it.
A bedroom that has a lot of clutter and a laptop on the bed.

High Caption Diversity (HCD)

Fixation Dataset: SALICON http://salicon.net/ Captioning Dataset: MSCOCO https://cocodataset.org/

多模注意机制 – 概念

XIAMEN UNIVERSITY

- 工作模式的挖掘
  - 描述一致性 CapSim (基于CIDEr定义)
  - 模式设置: HCD (低 CapSim) vs. HCC (高 CapSim)
  - 模式可预测性: VggNet, > 90% 准确率

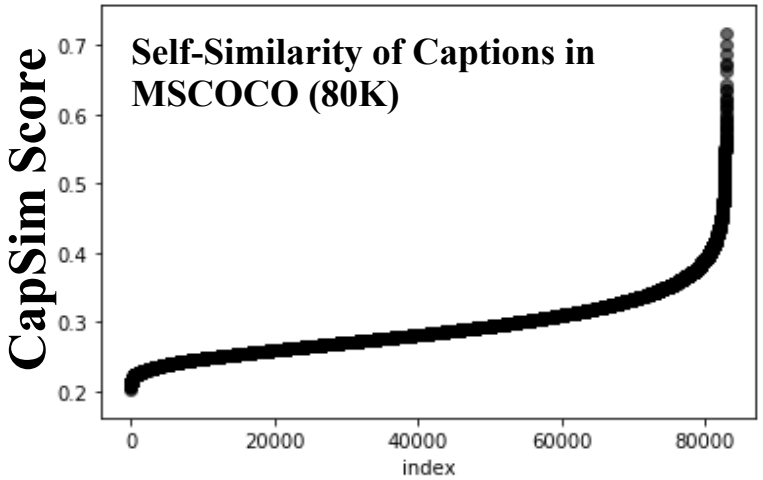
每张图片对应的captions数目

一张图片的任意两个caption

$$CapSim = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{q=1}^{T} CIDEr(c_t, c_q)$$

一个综合的caption评价指标

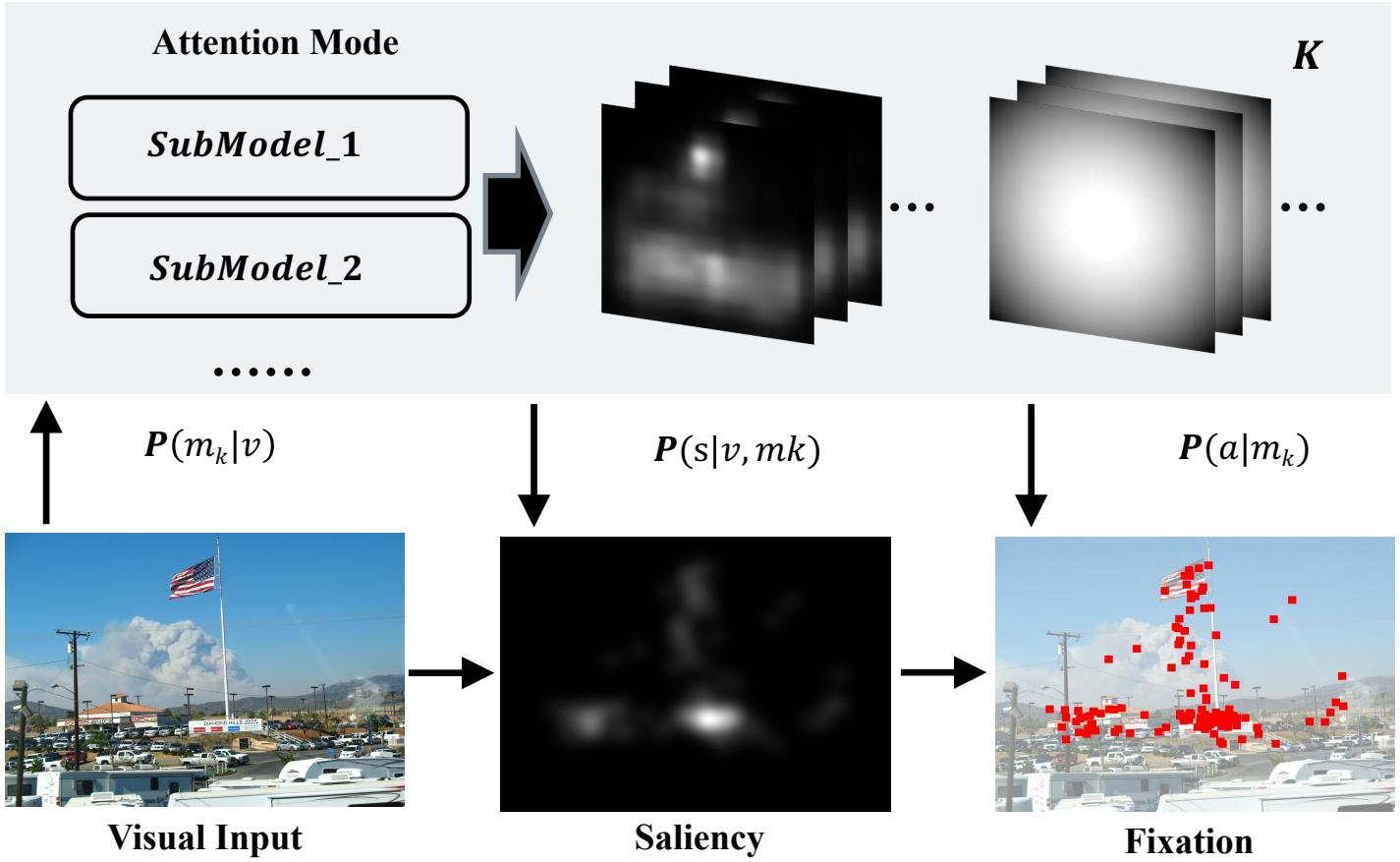


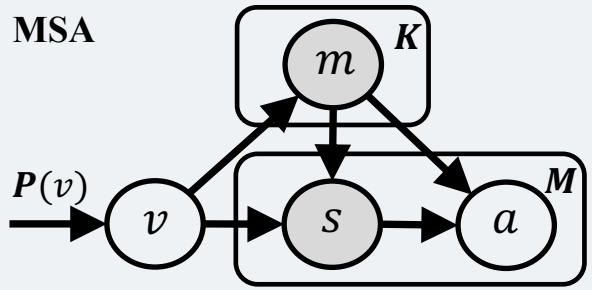


给定图片，单纯依靠视觉模型可基本准确预测该图片会产生一致性描述或是差异化描述

多模注意机制 – 方法

注意模型构建





$\boldsymbol{v}$: Visual input

$\boldsymbol{K}$: Number of Modes

$\boldsymbol{m}$: Attention mode

$\boldsymbol{s}$: Saliency

$\boldsymbol{M}$: Number of Locations

$\boldsymbol{a}$: Eye fixation

$$P(a_l|v_i) = \sum_{k=1}^{K} P(m_k|v_i)P(s_l|m_k,v_i)P(a_l|m_k)$$

$$P(a_l|m_k) = \mathcal{N}\left(log \sum_{d \in D_k} P_d(a_l)\right)$$

GitHub project page: https://github.com/zhangxuying1004/MSA

多模注意机制 – 实验

XIAMEN UNIVERSITY

实验结果

数据集: YORK120, MIT1003, PASCAL

指标：IG, IG-DGII, AUC, sAUC, NSS

Model	YORK120		MIT1003		PASCAL	
	$IG_{cen}\uparrow$	$IG_{dg2}\uparrow$	$IG_{cen}\uparrow$	$IG_{dg2}\uparrow$	$IG_{cen}\uparrow$	$IG_{dg2}\uparrow$
CEN [5]	0.000	-0.660	0.000	-0.939	0.000	-1.255
AWS [19]	0.358	-0.302	0.261	-0.678	0.242	-1.012
SAL [10]	0.927	0.267	1.050	0.111	1.230	-0.025
DGII [17]	0.660	0.000	0.939	0.000	**1.254**	**0.000**
MSA-S (Ours)	**1.015**	**0.355**	**1.061**	**0.122**	1.218	-0.036
MSA-A (Ours)	**1.025**	**0.365**	**1.083**	**0.144**	**1.257**	**0.003**

Model	YORK120			MIT1003			PASCAL		
	AUC↑	sAUC ↑	NSS↑	AUC↑	sAUC↑	NSS↑	AUC↑	sAUC↑	NSS↑
CEN [5]	0.784	0.500	0.959	0.787	0.513	0.968	0.840	0.516	1.159
AWS [19]	0.760	**0.717**	1.236	0.736	0.686	1.075	0.747	0.657	1.122
SAL [10]	0.819	0.707	1.904	0.830	0.719	1.976	0.854	**0.717**	2.210
DGII [17]	0.783	**0.717**	**2.280**	0.770	**0.749**	**2.681**	0.825	0.700	**2.546**
MSA-S (Ours)	**0.828**	0.708	1.928	**0.839**	**0.723**	1.978	**0.857**	0.712	2.205
MSA-A (Ours)	**0.828**	0.702	**1.951**	**0.839**	0.717	**1.998**	**0.858**	**0.713**	**2.235**

MSA Alternatives	YORK120		MIT1003		PASCAL		Mode Number
	IG_{cen}	IG_{dg2}	IG_{cen}	IG_{dg2}	IG_{cen}	IG_{dg2}	
MSA-S $_{HCC+HCD}$	1.015	0.355	1.061	0.122	1.218	-0.036	2
MSA-S $_{Topics,\ 5}$	1.031	0.370	1.060	0.121	1.239	-0.015	5
MSA-S $_{Topics,\ 10}$	1.036	0.376	1.066	0.127	1.220	-0.034	10
MSA-A $_{HCC+HCD}$	1.025	0.365	1.083	0.144	1.257	0.003	2
MSA-A $_{Topics,\ 5}$	1.032	0.372	1.083	0.144	1.273	0.018	5
MSA-A $_{Topics,\ 10}$	1.049	0.389	1.086	0.147	1.248	-0.006	10

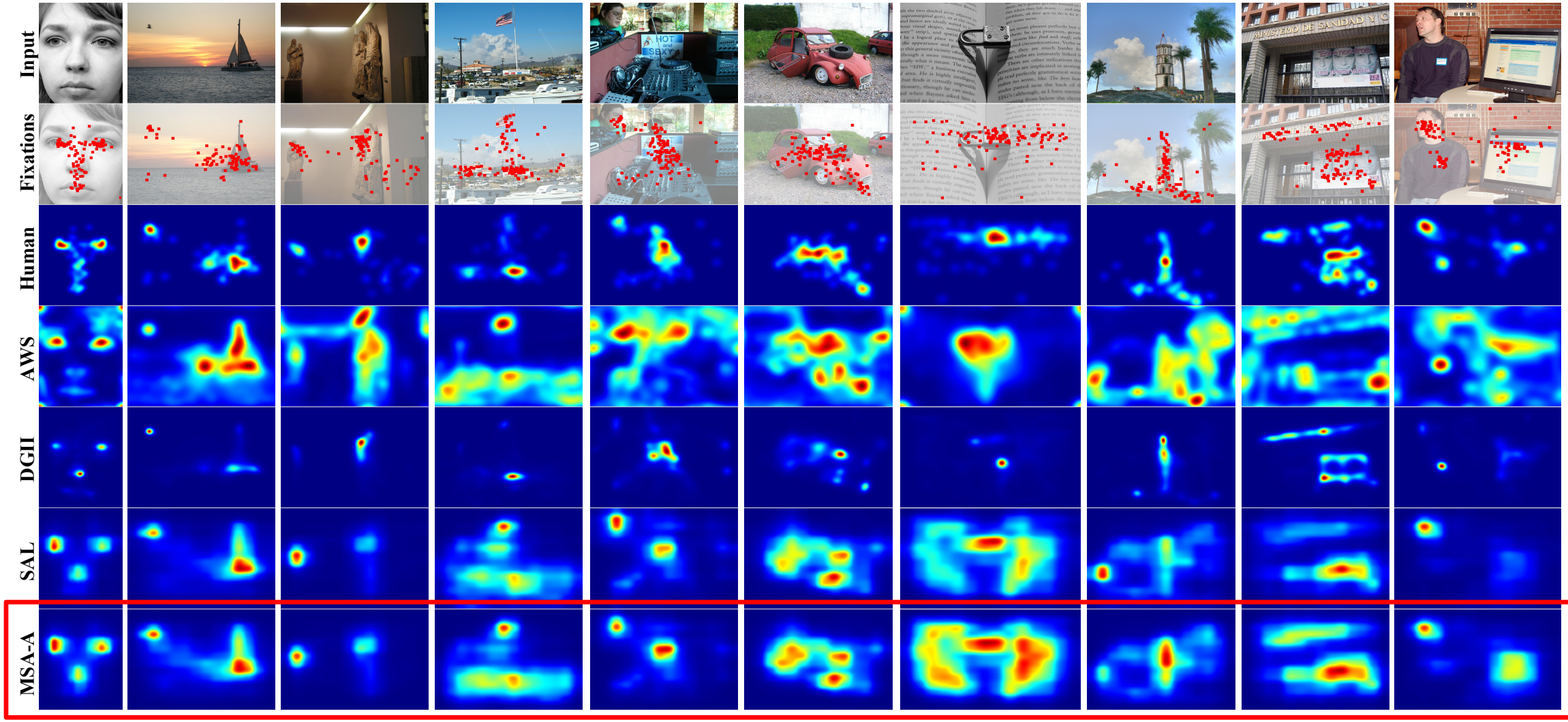
I) 以SAL模型为基线，采用MSA框架的模型性能在各项指标上均有明显提升

II) 利用语言描述和K均值聚类定义N个Topic模式，可观测到进一步提升，但尚未发现明确规律

多模注意机制 – 实验

XIAMEN UNIVERSITY

实验结果



04 多特征-多任务协同

Multi-Feature and Multi-task Collaboration





Yunpeng Luo

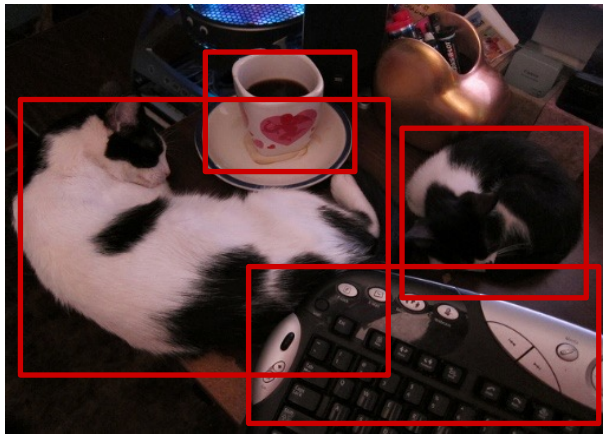
Dual-Level Collaborative Transformer for Image Captioning. AAAI 2021

多特征协同

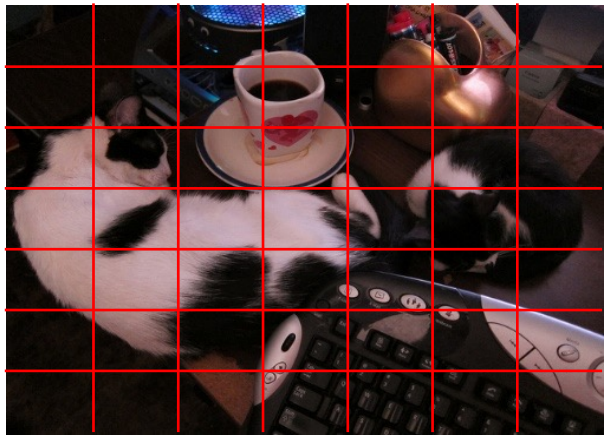
多特征协同 – 概念

XIAMEN UNIVERSITY

- 基于双层级特征协作的图像描述
  - 区域特征和网格特征各自存在局限性，可以通过两者互补解决
  - 特征交互时会出现语义噪声问题，导致性能提升不明显，甚至下降



区域特征关注目标信息，缺少细节和全局信息



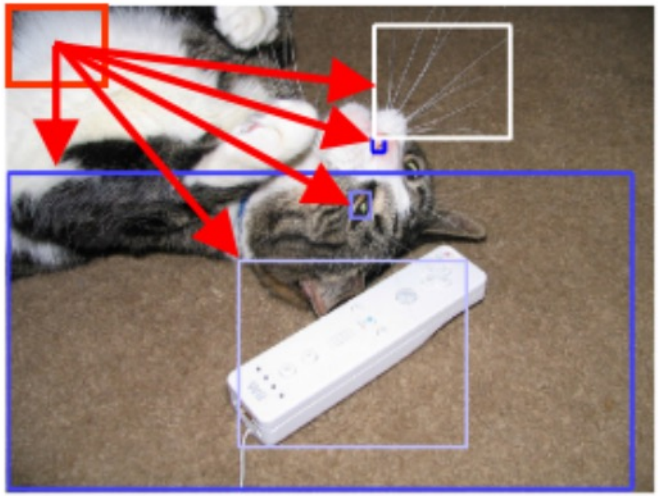
网格特征关注细节和全局信息信息，缺少语义层级较高的目标信息

特征	目标信息	细节和全局信息
区域特征	√	×
网格特征	×	√
区域特征 + 网格特征	√	√

多特征协同 – 概念

XIAMEN UNIVERSITY

- 基于双层级特征协作的图像描述
  - 区域特征和网格特征各自存在局限性，可以通过两者互补解决
  - 特征交互时会出现语义噪声问题，导致性能提升不明显，甚至下降



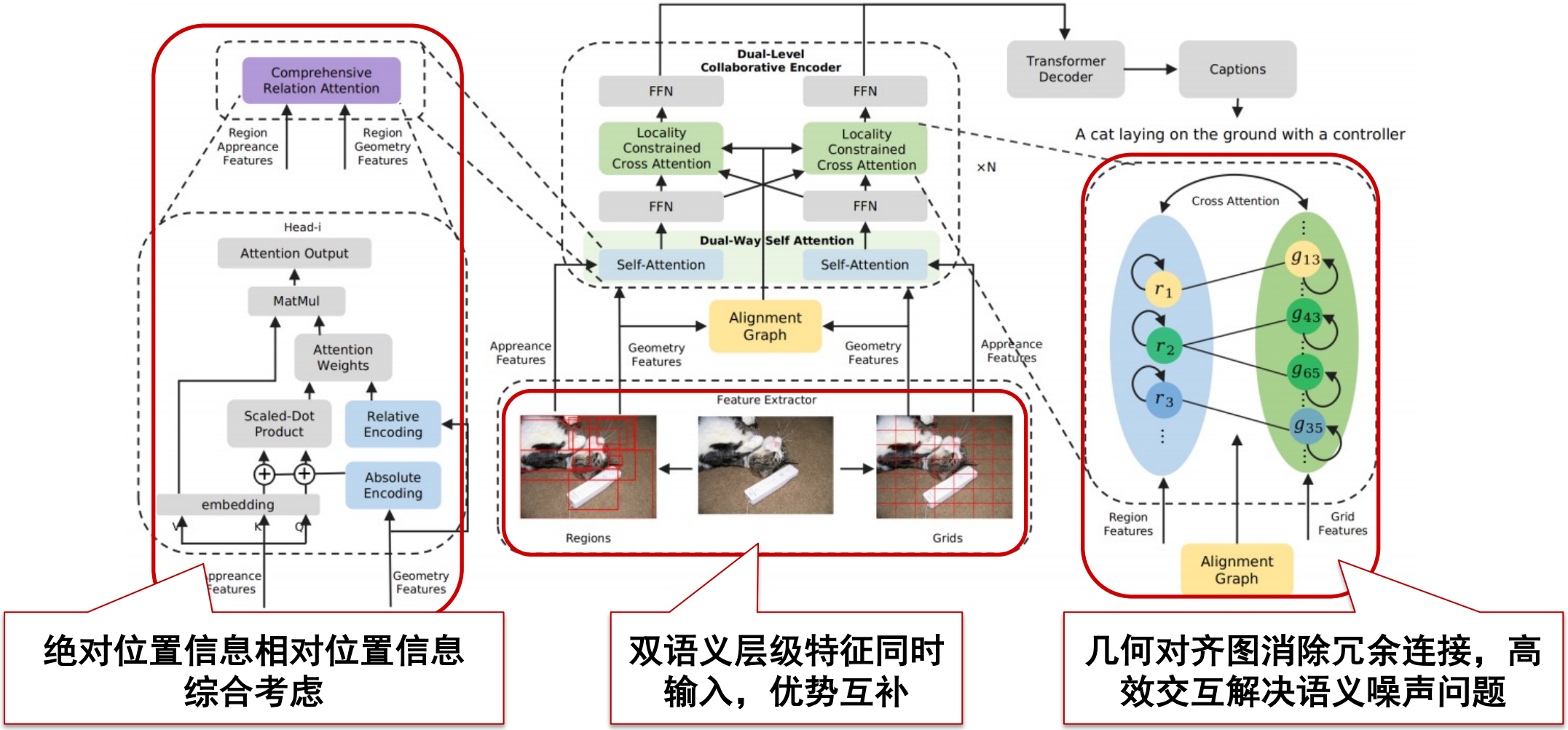
模型会关注到完全无关的区域，导致引入语义噪声

	B-1	B-4	M	R	C	S
Grid (G)	81.2	39.0	29.0	58.6	131.2	22.4
Region (R)	80.1	39.0	28.9	58.6	130.1	22.4
G + R	80.9	38.9	29.2	58.6	131.6	22.7
DLCT (G+R)	**81.4**	**39.8**	**29.5**	**59.1**	**133.8**	**23.0**
DLCT w/o LCCA	81.2	39.2	29.2	58.6	132.6	22.8
LCCA + CBG	80.8	38.7	29.0	58.7	130.8	22.7
DLCT	**81.4**	**39.8**	**29.5**	**59.1**	**133.8**	**23.0**

直接合并特征带来的提升并不明显,在不加限制的特征交互中甚至会导致效果下降

多特征协同 – 方法

模型构建



绝对位置信息相对位置信息综合考虑

双语义层级特征同时输入，优势互补

几何对齐图消除冗余连接，高效交互解决语义噪声问题

多特征协同 – 实验

XIAMEN UNIVERSITY

实验结果

Model	Single Model						Ensemble Model					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
SCST (ResNet-101) cvpr2017	-	34.2	26.7	57.7	114.0	-	-	35.4	27.1	56.6	117.5	-
Up-Down (ResNet-101) cvpr2018	79.8	36.3	27.7	56.9	120.1	21.4	-	-	-	-	-	-
HAN (ResNet-101) aaai2019	80.9	37.6	27.8	58.1	121.7	21.5	-	-	-	-	-	-
GCN-LSTM (ResNet-101) eccv2018	80.5	38.2	28.5	58.5	128.3	22.0	80.9	38.3	28.6	58.5	128.7	22.1
SGAE (ResNet-101) cvpr2019	80.8	38.4	28.4	58.6	127.8	22.1	81.1	39.0	28.4	58.9	129.1	22.2
ORT (ResNet-101) nips2019	80.5	38.6	28.7	58.4	127.8	22.1	-	-	-	-	-	-
SRT (ResNet-101) aaai2020	80.3	38.5	28.7	58.4	129.1	22.4	-	-	-	-	-	-
AoA (ResNet-101) iccv2019	80.2	38.9	29.2	58.8	129.8	22.4	81.6	40.2	29.3	59.4	132.0	22.8
AoA (ResNeXt-101 Grid) iccv2019	80.7	39.0	28.9	58.7	129.5	22.6	-	-	-	-	-	-
HIP (SENet-154) iccv2019	-	39.1	28.9	59.2	130.6	22.3	-	-	-	-	-	-
M2 (ResNet-101) cvpr2020	80.8	39.1	29.2	58.6	131.2	22.6	82.0	40.5	29.7	59.5	134.5	23.5
M2 (ResNeXt-101 Region) cvpr2020	80.6	38.8	29.0	58.4	130.8	22.4	-	-	-	-	-	-
M2 (ResNeXt-101 Grid) cvpr2020	80.8	38.9	29.1	58.5	131.7	22.6	-	-	-	-	-	-
X-Transformer (ResNet-101) cvpr2020	80.9	39.7	**29.5**	**59.1**	132.8	**23.4**	81.7	40.7	**29.9**	59.7	135.3	**23.8**
X-Transformer (ResNeXt-101 Grid) cvpr2020	81.0	39.7	29.4	58.9	132.5	23.1	-	-	-	-	-	-
Ours (ResNeXt-101)	**81.4**	**39.8**	**29.5**	**59.1**	**133.8**	23.0	**82.2**	**40.8**	**29.9**	**59.8**	**137.5**	23.3

#	User	Entries	Date of Last Entry	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr-D	
				c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲	c5 ▲	c40 ▲
1	yun3300612	2	09/03/20	0.824 (1)	0.966 (1)	0.674 (1)	0.917 (1)	0.528 (2)	0.838 (1)	0.406 (2)	0.740 (1)	0.298 (2)	0.396 (1)	0.598 (2)	0.753 (3)	1.333 (1)	1.354 (1)
2	IVA-CASIA	1	06/14/20	0.822 (2)	0.961 (4)	0.673 (2)	0.912 (3)	0.529 (1)	0.833 (2)	0.408 (1)	0.734 (2)	0.298 (3)	0.394 (3)	0.599 (1)	0.755 (2)	1.320 (2)	1.342 (2)
3	Prophet	1	06/02/20	0.818 (7)	0.963 (2)	0.665 (8)	0.912 (2)	0.519 (8)	0.832 (3)	0.398 (10)	0.733 (3)	0.296 (5)	0.393 (4)	0.594 (7)	0.751 (4)	1.304 (5)	1.337 (3)

MSCOCO线上和线下测试均达到当时最佳




Gen Luo Yiyi Zhou

Multi-task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. CVPR 2020

跨任务协同

跨任务协同 – 概念

XIAMEN UNIVERSITY

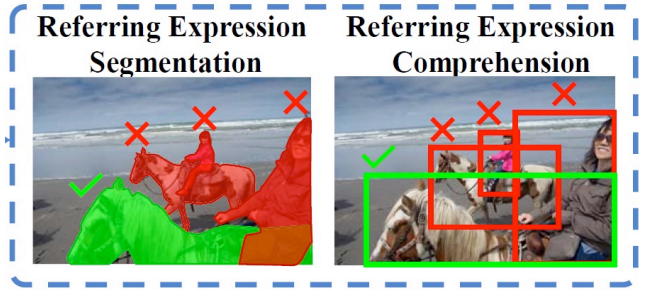
- 指向性检测与指向性分割
  - 旨在根据自然语言指令来检测/分割其语义指向的物体
  - 现有方法大多基于多阶段的方法，速度慢
  - 检测和分割两个任务间在单阶段网络下能够相互促进
  - 复杂场景下会出现不同任务的预测分歧问题
- 前人工作
  - MAttNet (Yu et al. CVPR 2018)
  - LGRANs (Wang et al. CVPR 2019)
  - NMTree (Liu et al. ICCV 2019)
  - FAOA (Yang et al. ICCV 2019)
  - CMSA (Ye et al. CVPR 2019)
  - ……

多阶段多任务网络
速度非常慢，受限于视觉预训练

单阶段单任务网络
速度快但精度低，无法执行多个任务



“a half horse.”





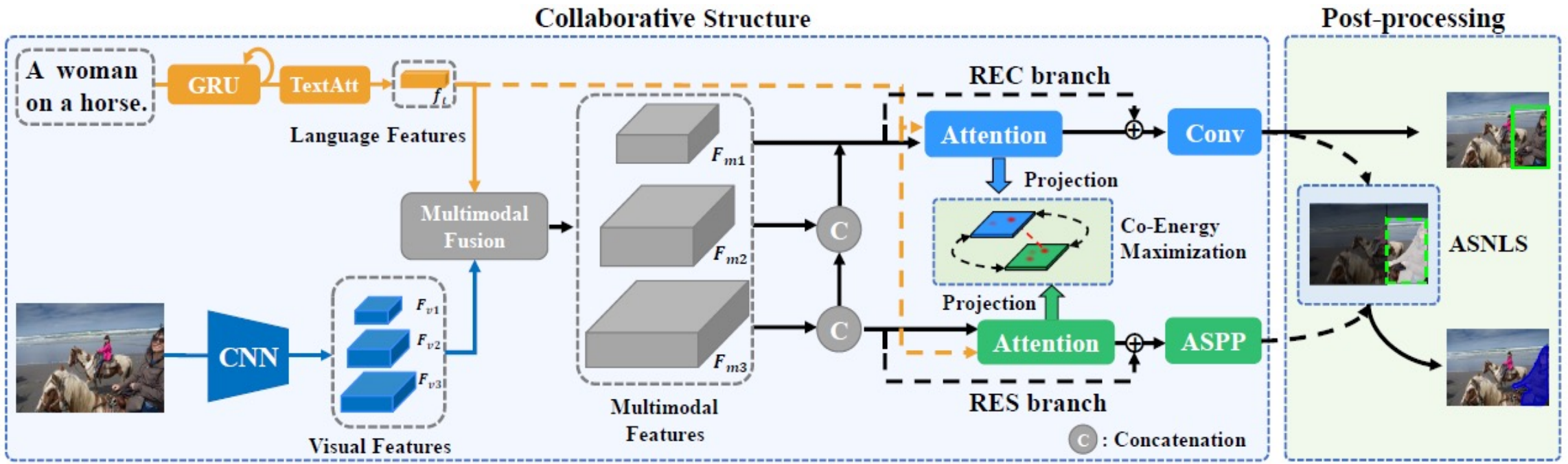
“person on scooter wearing black helmet and has black backpack”

跨任务协同 – 方法

XIAMEN UNIVERSITY

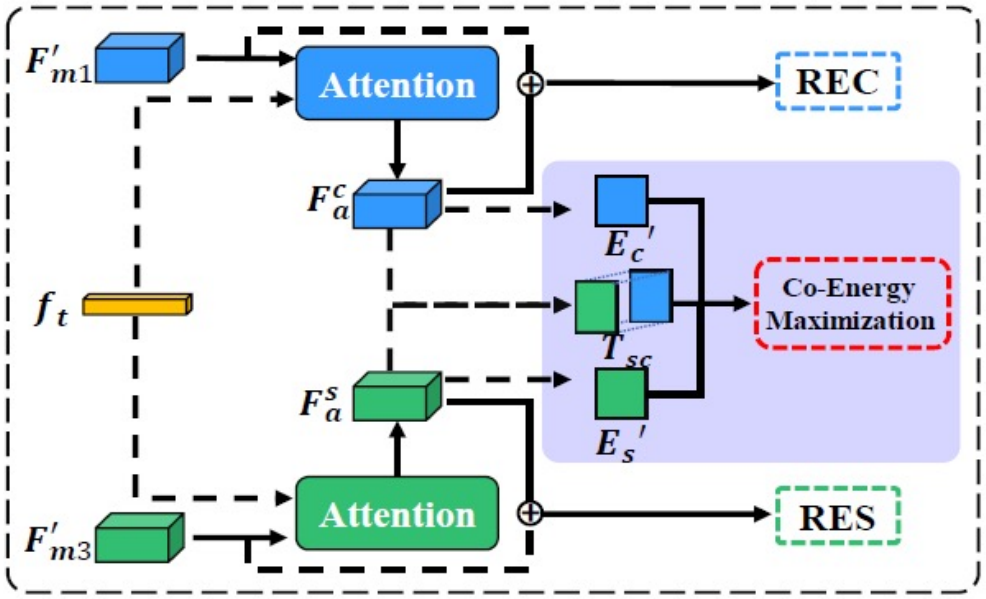
模型构建

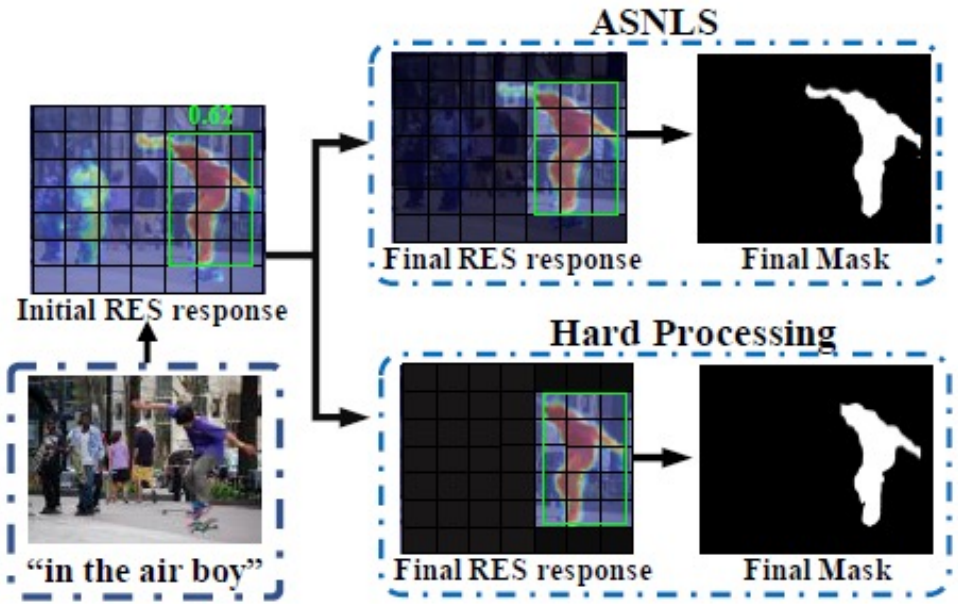
- 模型：多任务协同的指向性目标检测-分割网络
- 特性：解决多任务下的预测分歧问题，在双任务上均显著超越过去方法



跨任务协同 – 方法

- 多任务协同机制
  - 协同能量最大化模块
  - 自适应的非极大值抑制





- ✓ 最大化任务间的一致性
- ✓ 跨任务知识蒸馏

- ✓ 增强任务间的一致性
- ✓ 增强分割的鲁棒性

跨任务协同 – 实验

XIAMEN UNIVERSITY

实验结果

Table 4. Comparisons of MCN with the state-of-the-arts on the REC task.

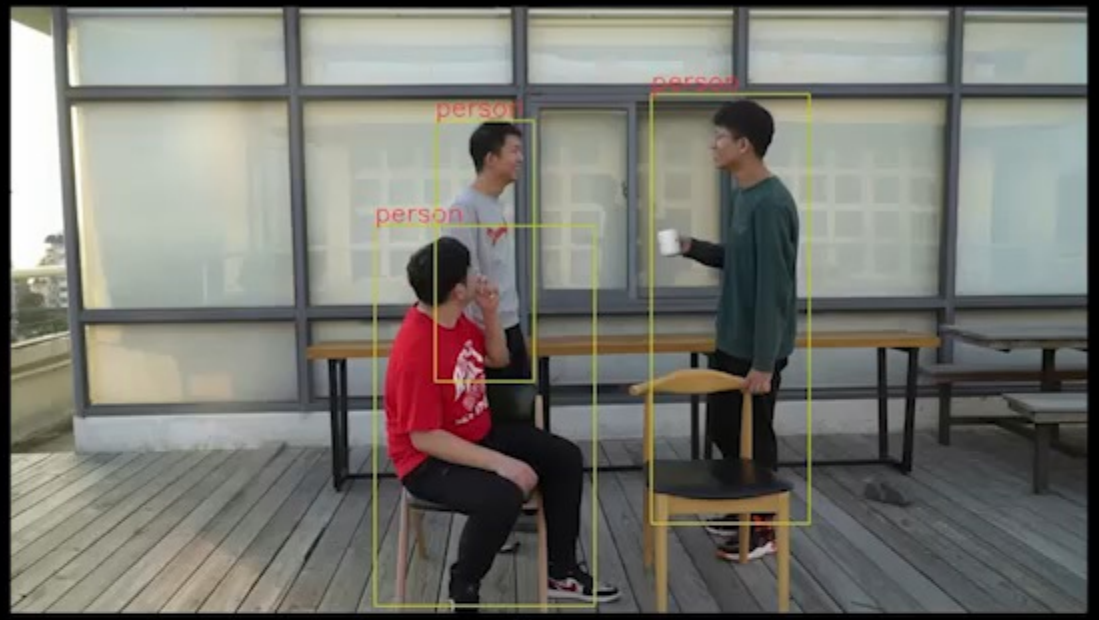
Model	Visual Features	RefCOCO			RefCOCO+			RefCOCOg		Speed* ↓
		val	testA	testB	val	testA	testB	val	test	
MMI [23] CVPR16	vgg16	-	64.90	54.51	-	54.03	42.81	-	-	-
CMN [31] CVPR16	vgg16	-	71.03	65.77	-	54.32	47.76	-	-	-
Spe+Lis+RI [45] CVPR17	frcnn-resnet101	69.48	73.71	64.96	55.71	60.74	48.80	60.21	59.63	-
Spe+Lis+RI [45] CVPR17	frcnn-resnet101	68.95	73.10	64.85	54.89	60.04	49.56	59.33	59.21	-
ParalAttn [49] CVPR18	frcnn-vgg16	-	75.31	65.52	-	61.34	50.86	-	-	-
LGRANs [37] CVPR19	frcnn-vgg16	-	76.60	66.40	-	64.00	53.40	-	-	-
NMTree [18] ICCV19	frcnn-vgg16	71.65	74.81	67.34	58.00	61.09	53.45	61.01	61.46	-
FAOA [38] ICCV19	darknet53	71.15	74.88	66.32	56.86	61.89	49.46	59.44	58.90	**39 ms**
MattNet [43] CVPR18	frcnn-resnet101	76.40	80.43	69.28	64.93	70.26	56.00	**66.67**	67.01	367 ms
MattNet [43] CVPR18	mrcnn-resnet101	76.65	81.14	69.99	65.33	71.62	56.02	66.58	**67.27**	378 ms
MCN (ours)	vgg16	75.98	76.97	73.09	62.80	65.24	54.26	62.42	62.29	48 ms
MCN (ours)	darknet53	**80.08**	**82.29**	**74.98**	**67.16**	**72.86**	**57.31**	66.46	66.01	56 ms

* The inference time is tested on the same hardware, i.e., GTX1080ti.

Table 5. Comparisons of MCN with the state-of-the-arts on the RES task.

Model	Visual Features	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
DMN [25] ECCV18	resnet101	49.78	54.83	45.13	38.88	44.22	32.29	-	-
RRN [16] CVPR18	resnet101	55.33	57.26	53.93	39.75	42.15	36.11	-	-
CMSA [40] CVPR19	resnet101	58.32	60.61	55.09	43.76	47.60	37.89	-	-
MattNet [43] CVPR18	mrcnn-resnet101	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree [18] ICCV19	mrcnn-resnet101	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
MCN (ours)	vgg16	57.33	58.59	57.23	46.53	48.68	41.93	46.95	47.20
MCN (ours)	darknet53	**62.44**	**64.20**	**59.71**	**50.62**	**54.99**	**44.69**	**49.22**	**49.40**

MCN 超过多阶段模型的精度，同时保持了接近实时的推理速度



yolov3



RealGin



RealGin



RealGin

05 可控性
Controllability





Yiyi Zhou

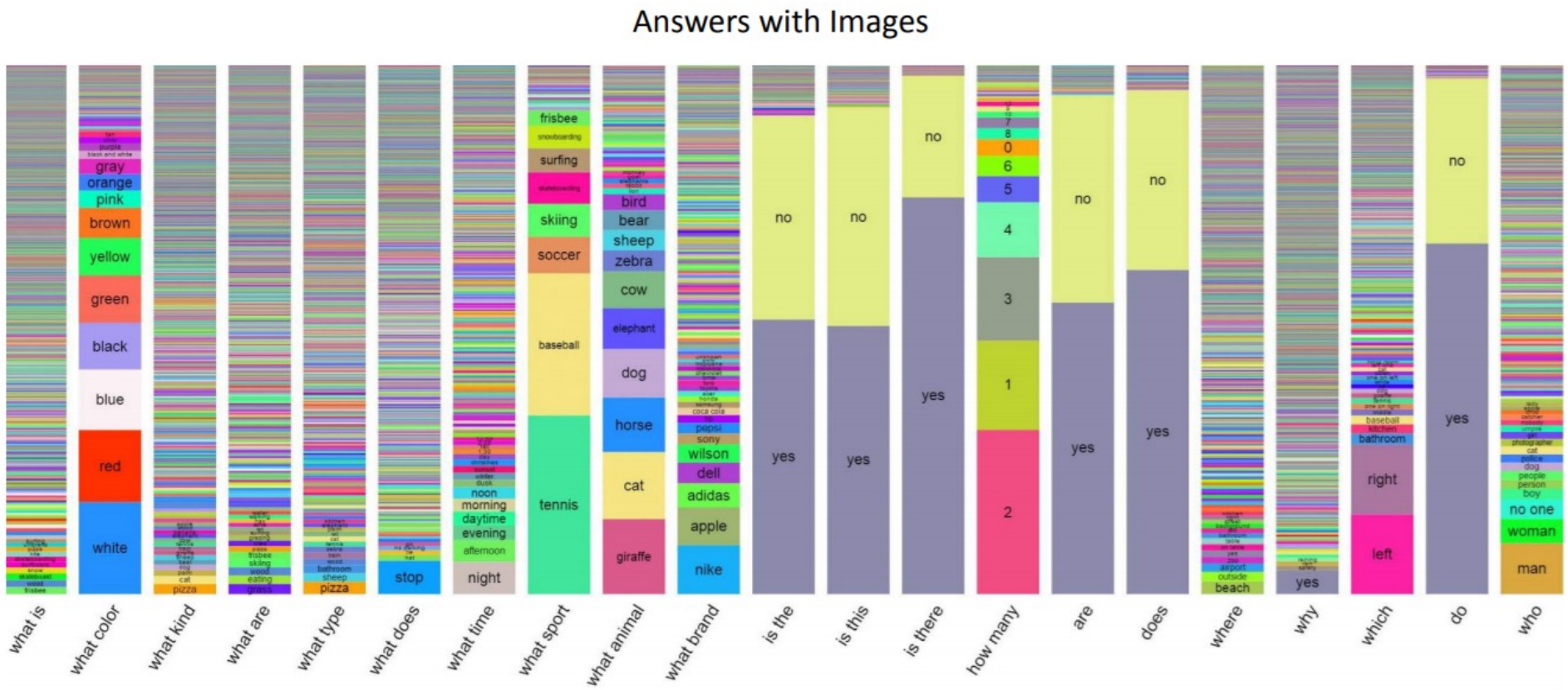
Free VQA Models from Knowledge Inertia by Pairwise Inconformity Learning, AAAI’19

训练可控性

训练可控性 – 概念

XIAMEN UNIVERSITY

- 多模态数据分布极度不均衡



VQA1.0 数据集的答案分布

当模型的训练和测试分布不一致时，模型性能显著下降

训练可控性 – 概念

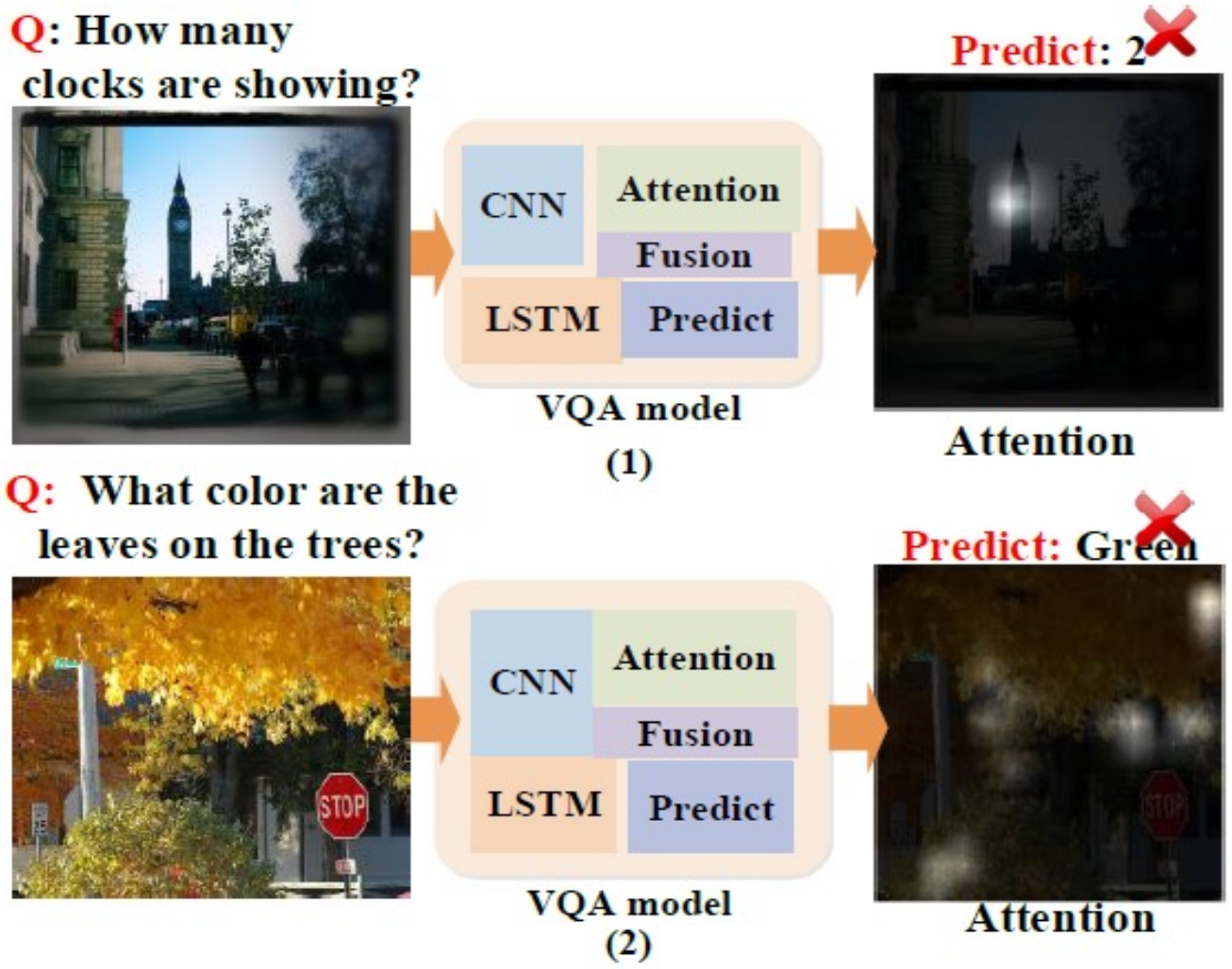
XIAMEN UNIVERSITY

- 强文本先验现象对模型预测产生重大影响
  - 视觉问答（Visual Question Answering）
    - 依赖问题的答案先验
  - 看图说话（Image Captioning）
    - 生成的文本风格单一、不符合人类描述习惯
- 模型的知识惯性
  - 知识惯性（Knowledge Inertia）
    - 使用旧的知识及过往经验来解决问题，而没有去探索新的信息及知识
  - 视觉问答中
    - 依靠问题描述去猜答案
    - 忽略了所给的视觉信息

训练可控性 – 概念

XIAMEN UNIVERSITY

- 模型的知识惯性与认知适应性相矛盾



2是How many问题中最常见的答案

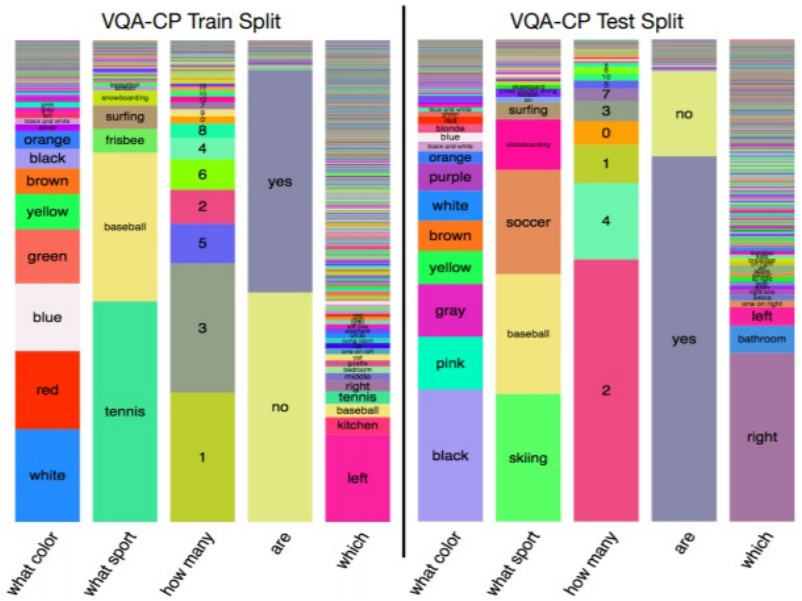
模型都关注到了正确的答案区域，但仍然回答错误

Green是数据集中学到的知识

模型根据训练的经验以及答案的偏执分布来预测答案称为知识惯性

训练可控性 – 概念

- 解决知识惯性的思路
  - 重新组织数据集分布
    - VQA2.0, VQA-CP, CLEVER
  - 提供正负样本对设置
    - VQA2.0 相似的图片、完全不同的答案



VQA-CP数据集训练和测试分布



VQA2.0数据集的相似样本对设置

训练可控性 – 方法

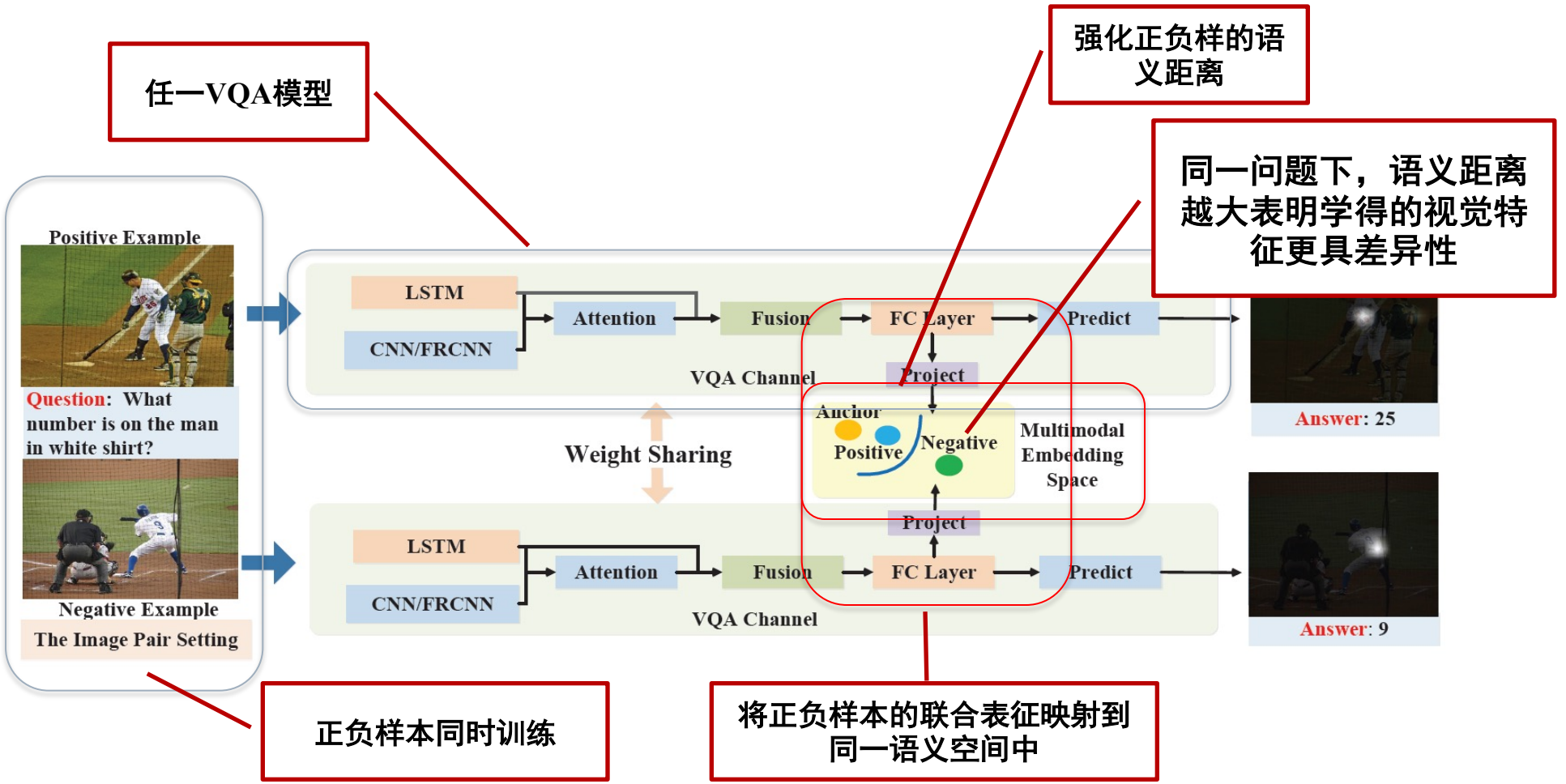
- 非对称学习策略（Pairwise Inconformity Learning, PIL）
  - 利用VQA2.0所提供的相似样本对标记
  - 通过相似样本对的联合学习，强迫模型更多地关注视觉内容
  - 让模型根据所“看见”的信息来进行预测



采用VQA2.0数据集的相似样本对设置，这种相似样本对的设置在数据集中比较容易获取
可以通过检索比较的方式自动构建这种标记信息，不需要大量的人工参与

训练可控性 – 方法

模型架构



训练可控性 – 方法

XIAMEN UNIVERSITY

- 多模态嵌入空间
  - 对正负样本的联合特征进行映射

$$f_e^p = \sigma\left(\mathbf{W}_{emb} f_j^p\right), \quad f_e^n = \sigma\left(\mathbf{W}_{emb} f_j^n\right),$$

  - 最小化与锚点特征的距离

$$\mathcal{L}_{emb} = \left\| f_e^p - f_a^p \right\|_2^2 + \left\| f_e^n - f_a^n \right\|_2^2,$$

  - 要点：
    - 直接加距离约束会损害模型性能
    - 答案的预训练词向量 (word embedding） 作为锚点特征, *e.g.*, GLOVE
      - 避免锚点特征的计算
      - 准确反映正负样本直接的语义距离

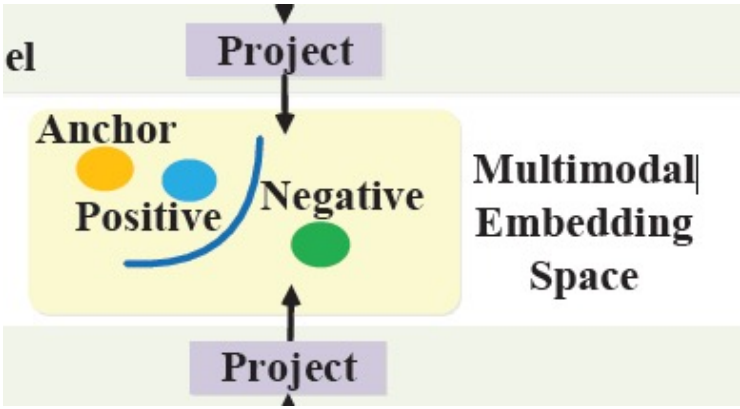
训练可控性 – 方法

■ 非对称学习策略（Pairwise Inconformity Learning）

□ 基于动态边际的三元组损失函数 (Dynamic-margin based Triple Loss)

$$\mathcal{L}_{triplet} = \max\left(0, m + \left\|f_e^p - f_a^p\right\|_2^2 - \left\|f_e^n - f_a^p\right\|_2^2\right)$$
$$- \max\left(0, m + \left\|f_e^n - f_a^n\right\|_2^2 - \left\|f_e^p - f_a^n\right\|_2^2\right),$$
$$where \quad m = \left\|f_a^p - f_a^n\right\|_2^2.$$

利用锚点避免边际调参，更好的实际语义边际反映



□ 困难样本挖掘

$$d_s = \begin{cases} 0.25 & if \quad d(f_e^p, f_e^n) > m, \\ 0.5 & if \quad d(f_e^p, f_a^p) < d(f_e^p, f_e^n) < m, \\ 1.0 & if \quad d(f_e^p, f_e^n) < d(f_e^p, f_a^p), \end{cases}$$

基于语义距离的困难样本挖掘

训练可控性 – 实验

XIAMEN UNIVERSITY

实验结果

VQA1.0 test-dev	All	Yes/No	Num.	Other
SAN[9]	58.7	79.3	36.6	46.1
DAMN[71]	59.4	81.1	38.6	45.4
HiCo[12]	61.8	79.7	38.7	51.7
MCB[10]	64.2	82.2	37.7	54.8
DAN$^{\mathcal{G}}$[108]	64.3	83.0	39.1	53.9
MLB$^{\mathcal{G}}$[59]	65.1	**84.1**	38.2	54.9
MFB[23]	**65.9**	84.0	**39.8**	**56.2**
Baseline (Ours)	64.7	83.2	37.9	54.9
Baseline_PIL (Ours)	66.1	83.7	39.4	56.7
Baseline$^{\mathcal{F}}$ (Ours)	66.9	**84.8**	40.8	57.5
Baseline_PIL$^{\mathcal{F}}$ (Ours)	**67.6**	83.7	**41.4**	**59.7**

VQA1.0 test	All	Yes/No	Num.	Other
SAN[9]	58.9	-	-	-
DAMN[71]	59.4	-	-	-
HiCo[12]	62.1	-	-	-
MCB$^{\mathcal{G}}$[10]	64.2	82.2	37.7	54.8
DAN$^{\mathcal{G}}$[108]	64.3	83.0	39.1	53.9
MLB$^{\mathcal{G}}$[59]	65.1	**84.1**	38.2	54.9
MFB$^{\mathcal{G}}$[23]	**65.9**	84.0	**39.8**	**56.2**
Baseline_PIL (Ours)	66.2	83.8	39.6	56.6
Baseline_PIL$^{\mathcal{F}}$ (Ours)	**67.7**	**83.7**	**41.3**	**59.6**

VQA2.0 test-dev	All	Yes/No	Num.	Other
MFB[23]	64.16	80.95	40.73	54.62
MF-SIG$^{\mathcal{G}}$[56]	64.73	81.29	42.99	55.55
BUA$^{\mathcal{G}}$[33]	62.07	79.20	39.46	52.62
BUA$^{\mathcal{F},\mathcal{G}}$[33]	**65.32**	**81.82**	**44.21**	**56.05**
Baseline (Ours)	63.53	80.40	41.16	54.23
Baseline_PIL (Ours)	65.07	81.15	43.20	56.34
Baseline$^{\mathcal{F}}$ (Ours)	66.21	82.38	45.03	57.24
Baseline_PIL$^{\mathcal{F}}$ (Ours)	**67.53**	**83.96**	**46.50**	**58.29**

VQA2.0 test	All	Yes/No	Num.	Other
LSTM-Q[18]	54.22	73.46	35.18	41.83
MCB[10]	62.27	78.82	38.28	53.36
MF-SIG$^{\mathcal{G}}$[56]	65.84	81.85	43.64	57.07
BUA$^{\mathcal{G}}$[33]	62.27	79.32	39.77	52.59
BUA$^{\mathcal{F},\mathcal{G}}$[33]	**65.67**	**82.20**	**43.90**	**56.26**
Baseline_PIL (Ours)	65.10	81.38	43.33	56.24
Baseline_PIL$^{\mathcal{F}}$ (Ours)	**67.65**	**83.91**	**46.01**	**58.46**

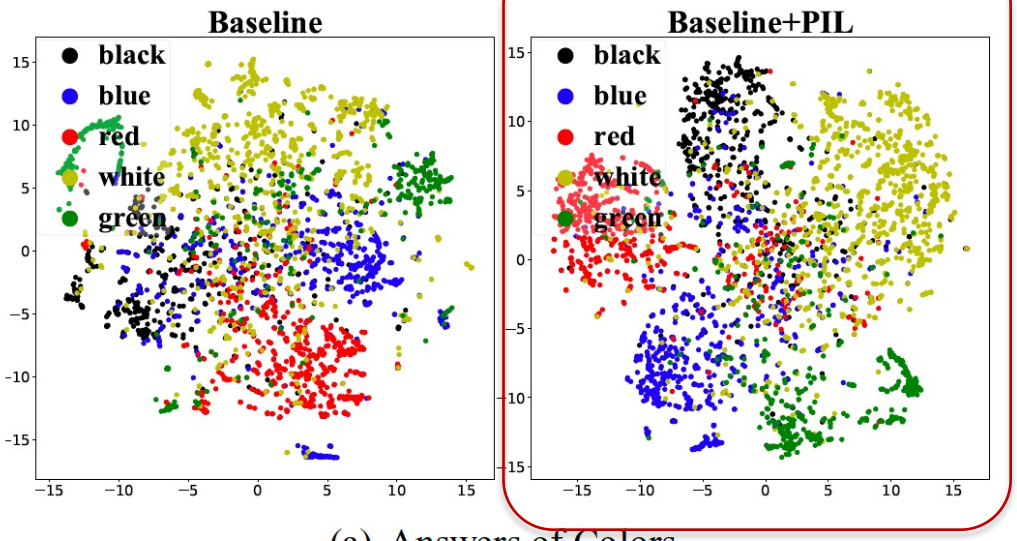
将一个简单的VQA模型提升到SOTA

训练可控性 – 实验

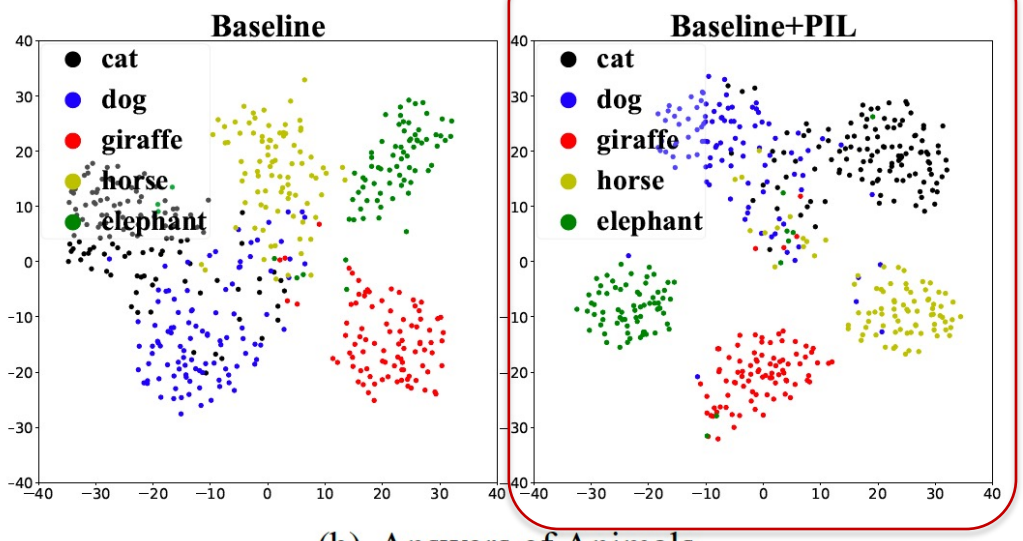
XIAMEN UNIVERSITY

实验结果

PIL让样本的特征更加具有判别性



(a) Answers of Colors



(b) Answers of Animals

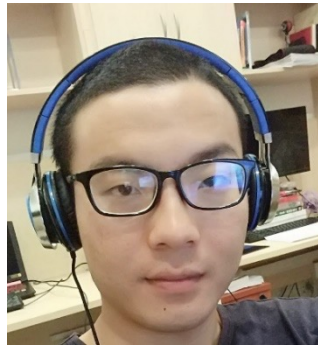
训练可控性 – 实验

XIAMEN UNIVERSITY

实验结果

所答即所“见”







Jiayi Ji　　Yiyi Zhou

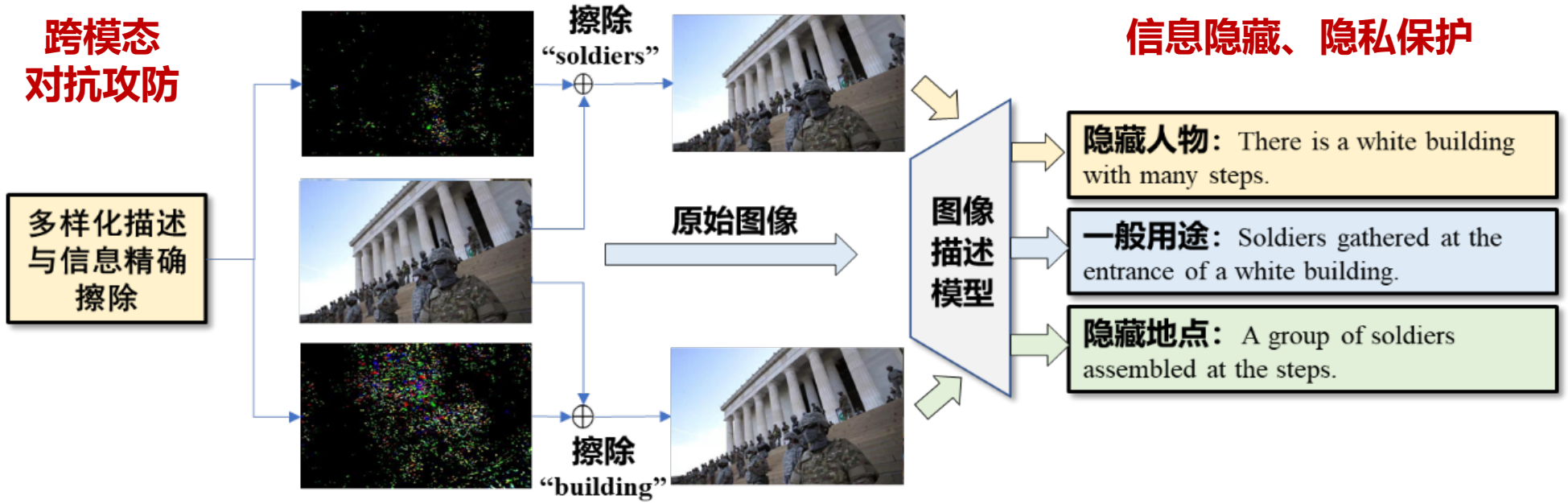
Attacking Image Captioning Towards Accuracy-Preserving Target Words Removal. ACM MM 2020

模型可控性

模型可控性 – 概念

XIAMEN UNIVERSITY

- 针对描述模型的目标语义擦除
  - 实现精准的语义攻击，即攻击特定的概念/词的同时，维持描述的准确性
  - 要求：图像改变人肉眼不可见；即插即用，仅针对测试阶段，与训练无关
  - 应用：跨模态攻防、多样化描述、隐私保护、设施安全

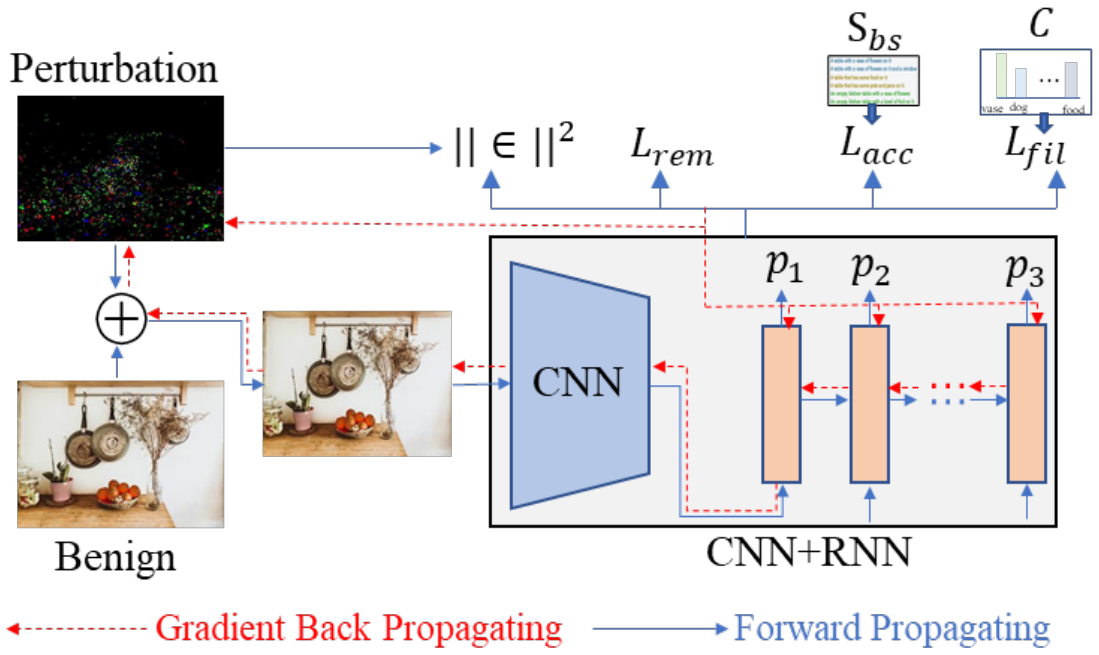


在图像中加入微弱干扰，在确保描述精确的前提下屏蔽特定概念词

模型可控性 – 方法

■ 针对描述模型的目标语义擦除

□核心：三个约束项



面向图像描述目标词移除的图像语义攻击

目标词移除约束

$$L_{rem} = \sum_{j=1}^{M} \max_{t\in\{1,2,\cdots,N\}} \left\{ \max \left\{ -\alpha, log(p_t^{D_j}) - log(p_t^{V^k}) \right\} \right\},$$

准确性保持约束

$$L_{acc} = -E_{\omega^s \sim P(S|D_{tar},\epsilon)}[r(\omega^s)],$$

有害噪声过滤约束

$$L_{rem} = \sum_{j=1}^{M} \max_{t\in\{1,2,\cdots,N\}} \left\{ \max \left\{ -\alpha, log(p_t^{D_j}) - log(p_t^{V^k}) \right\} \right\},$$

模型可控性 – 实验

XIAMEN UNIVERSITY

- 针对描述模型的目标语义擦除
  - 数据集：MIC-COCO
  - 指　标：攻击成功率、准确保持率、攻击质量（精度保持+攻击成功）

Implemented Model	Methods	B-1	B-2	B-3	B-4	M	R	C	S	AR	SR	AQ
ST	BeamSearch(20)	59.4	38.8	25.2	16.2	18.5	41.6	59.7	12.1	68.5	84.2	76.4
	BeamSearch(100)	56.7	36.8	25.0	17.5	18.1	40.5	55.9	10.9	64.1	90.9	77.5
	Masking BeamSearch	66.2	47.8	33.1	22.6	21.4	48.2	67.1	13.7	77.0	**100.0**	88.5
	Ours	**69.8**	**52.2**	**38.2**	**28.0**	**23.4**	**51.0**	**84.8**	**14.1**	**97.3**	93.8	**95.6**
SAT	BeamSearch(20)	64.4	45.1	30.1	20.6	20.8	43.8	71.1	13.2	77.1	75.3	76.2
	BeamSearch(100)	62.2	42.9	29.4	20.5	20.1	43.2	68.4	12.5	74.2	88.3	81.2
	Masking BeamSearch	67.4	49.3	34.3	23.4	22.1	48.9	70.5	14.5	76.4	**100.0**	88.2
	Ours	**71.5**	**53.1**	**39.1**	**29.6**	**24.2**	**51.6**	**89.8**	**14.6**	**97.4**	91.9	**94.6**
Up-Down	BeamSearch(20)	66.4	46.1	33.8	27.2	23.1	48.7	84.3	16.8	69.5	74.4	71.9
	BeamSearch(100)	64.1	44.0	30.1	23.6	22.3	45.5	76.1	15.4	62.7	87.2	75.0
	Masking BeamSearch	69.9	52.3	38.5	29.2	23.6	51.3	91.3	19.7	75.3	**100.0**	87.7
	Ours	**74.5**	**63.5**	**47.6**	**35.1**	**26.7**	**55.3**	**118.8**	**20.3**	**97.9**	96.9	**97.4**

相比之下，我们的模型在准确率保持（AR）和攻击质量(AQ)上都大幅高于基线模型

Component	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE	AR	SR	AQ
ADJ	**75.3**	**57.4**	**42.1**	**30.4**	**25.3**	**53.1**	**91.5**	**14.9**	**99.2**	**98.2**	**98.7**
VERB	71.2	54.1	37.9	26.8	21.5	49.3	88.4	13.1	95.9	92.9	94.4
NOUN	71.6	52.7	39.1	28.0	23.4	51.0	89.2	14.3	96.7	92.1	94.4

名词、动词相较形容词更难被攻击

06 总结
Conclusions



总 结

XIAMEN UNIVERSITY

研究方面

- 语言的统计先验具有极高价值，尤其是概念的**层次结构、分布特性**
- **细粒度**和**高判别力**特征对保证跨模态任务精度至关重要
- 视觉-语言的研究应是**双向的**，各自的新发现、新方法可以互相迭代促进
- **跨空间、跨领域、跨任务**知识迁移方面，有较大探索空间，易于发现新的科学问题

应用方面

- 优化**训练策略**可解决样本分布不均衡，数据冗余等复杂问题
- 模型的**鲁棒性、可控性**是相关技术落地的最后一道坎，不容小觑

MAC实验室概况

XIAMEN UNIVERSITY

厦门大学媒体分析与计算实验室（MAC）https://mac.xmu.edu.cn/

- 纪荣嵘（国家杰青）教授创建，包括专任科研人员 11人



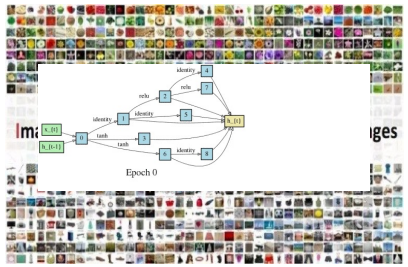
- 实验室分为四个研究组（研究生70余人）



媒体检索组



社交媒体组



机器学习组



图像理解组

- 媒体检索
- 人车物监控
- 多模态分析
- 舆情感知引导
- 自动机器学习
- 神经网络压缩
- AI模型鲁棒性
- 目标检测识别

谢谢

THANK YOU

https://mac.xmu.edu.cn/

