京东智联云

Vision and Language: from Perception to Creation

Tao Mei, Ph.D.

AI Research, JD.COM

tmei@jd.com

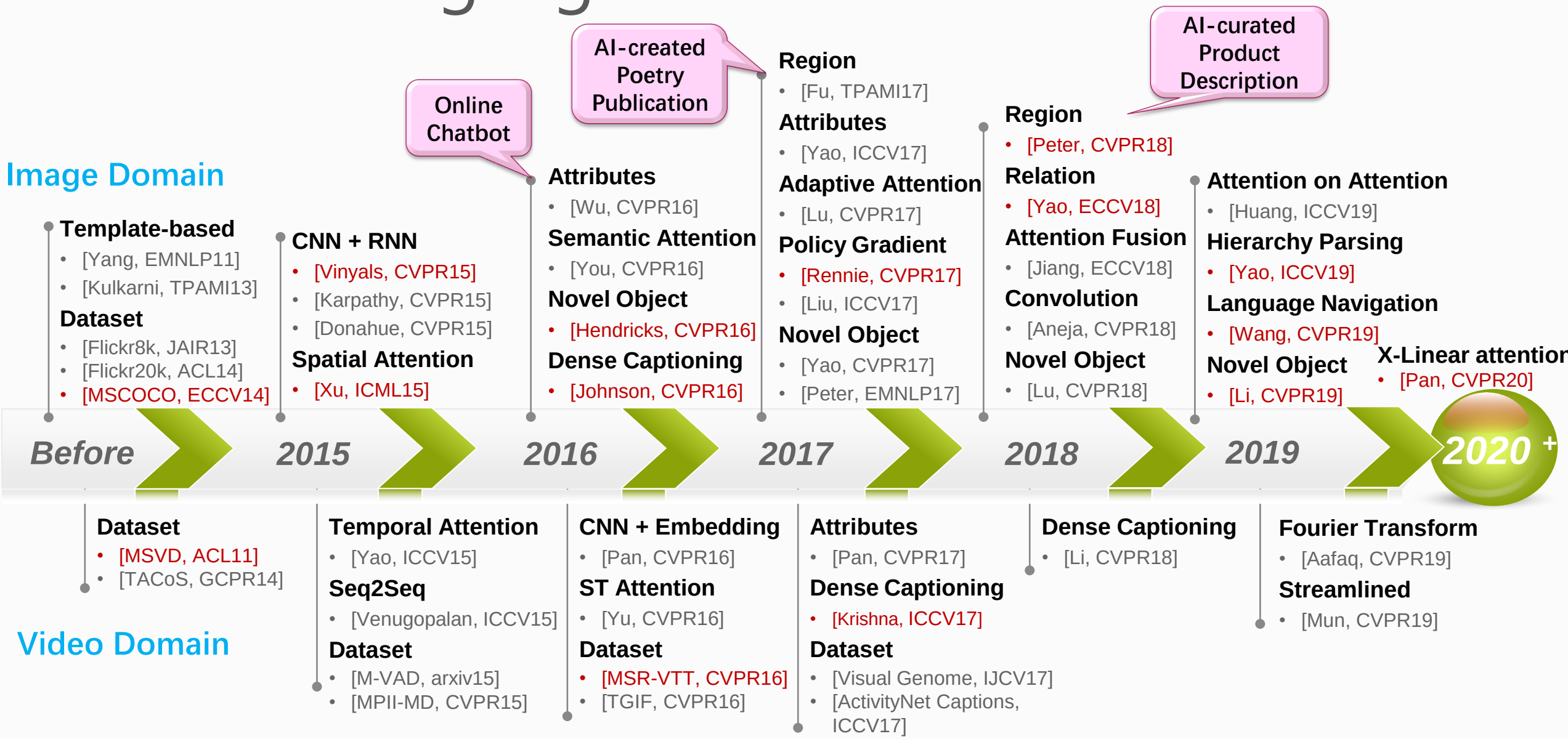
image captioning

a group of zebras standing in a field [Vinyals, CVPR15]
a group of zebras grazing on grass [You, CVPR16]
a group of zebras grazing in a field [Yao, ICCV17]
a group of zebras and a rainbow in the sky [Peter, CVPR18]
a group of zebras grazing in a field with a rainbow in the sky [Yao, ECCV18]
a group of zebras and other animals grazing in a field with a rainbow in the sky [Pan, CVPR20]

video captioning

a man is singing a song [Venugopalan, ICCV15]
a man is dancing [Pan, CVPR16]
a woman is dancing [Pan, CVPR17]
a woman is dancing in the rain [Chen, AAAI19]

Vision to Language



Mainstream: CNN Encoder + LSTM Decoder

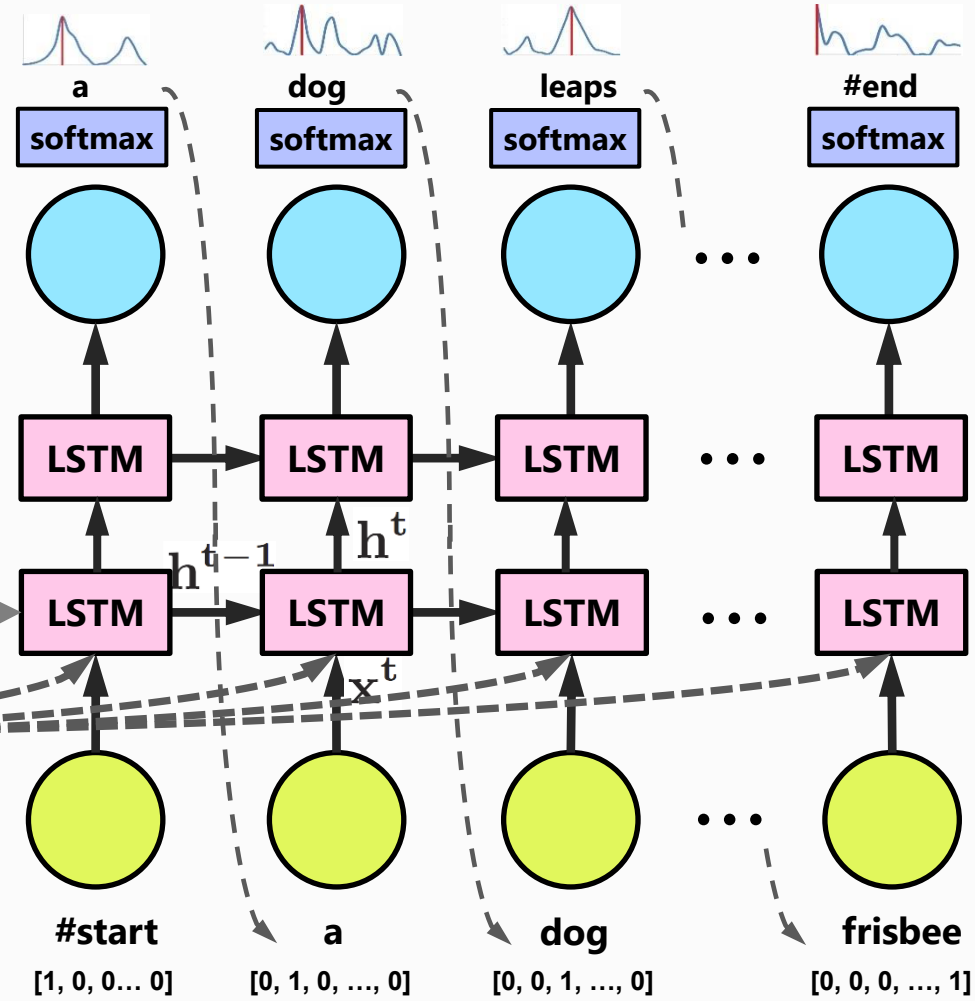
[Google15, Stanford15, Berkeley15, Baidu/UCLA15, UdeM15, Rochester16, UAdelaide16, Virginia Tech17, THU17, MSR17&18, IBM17, U of Oxford & Google17, JD AI18&19]

Learning visual representation by CNN

- CNN Rep. — [Vinyals, CVPR15; Karpathy, CVPR15; Donohue, CVPR15; Mao, ICLR15]
- Attributes — [Wu, CVPR16; Yao, ICCV17]
- Attention — [Xu, ICML15; You, CVPR16; Lu, CVPR17]
- Objects — [Yao, CVPR17]
- Region — [Fu, TPAMI17; Peter, CVPR18]
- Relation — [Yao, ECCV18]
- Hierarchy — [Yao, ICCV19]

Policy Gradient Optimization

[Rennie, CVPR17; Liu, ICCV17]



* Note that this figure only shows prediction process.

Technology Trends

- Aim for a thorough image/video understanding for captioning
  -> Direction I: enhance image encoder with X

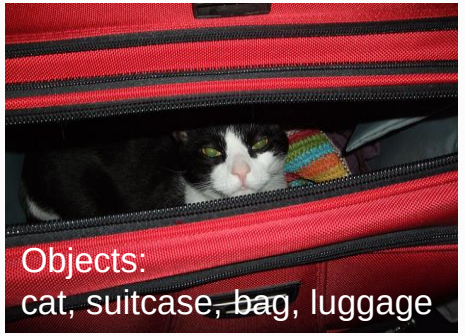
- Explore the (1^{st} , 2^{nd} , ...) interaction across multi-modal inputs
  (hidden states of sentence decoder & the encoded image features)
  -> Direction II: integrate encoder/decoder via X attention

- Learn a universal encoder-decoder structure for VL tasks
  -> Direction III: vision-language pre-training

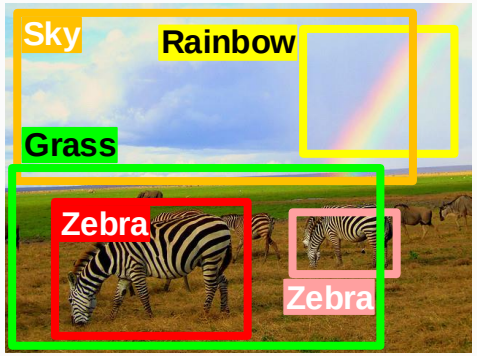
Direction I: enhance image encoder with X

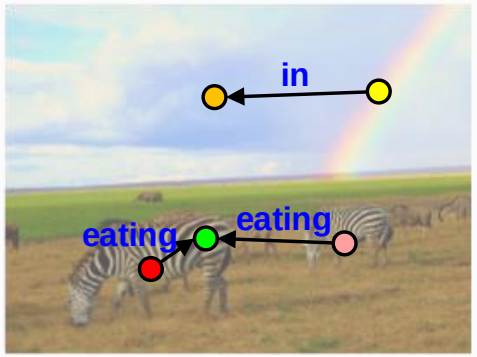
X = visual attributes
[You, CVPR16; Wu, CVPR16; **Yao, ICCV17**]

X = object/entity
[**Yao, CVPR17**; **Li, CVPR19**]

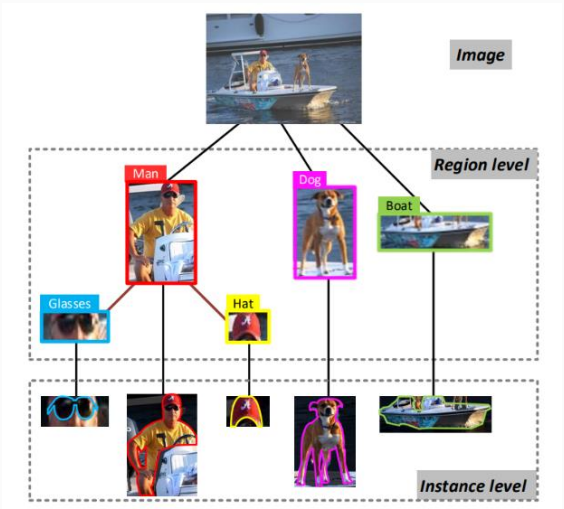


X = region/relation
[Peter, CVPR18; **Yao, ECCV18**]



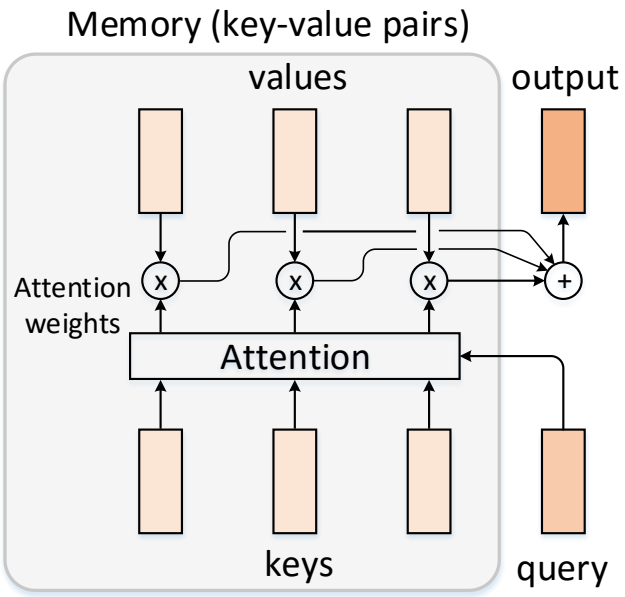


X = instance/hierarchy
[**Yao, ICCV19**]



Direction II: integrate encoder/decoder via X attention

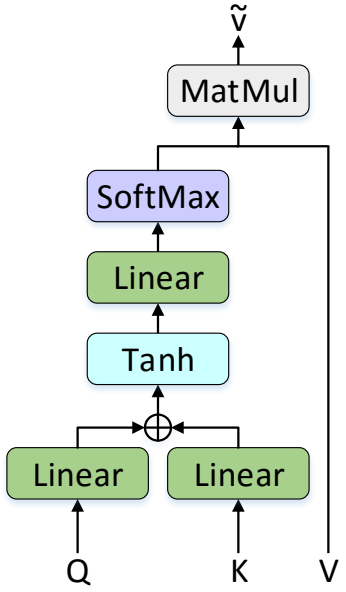
basic concept



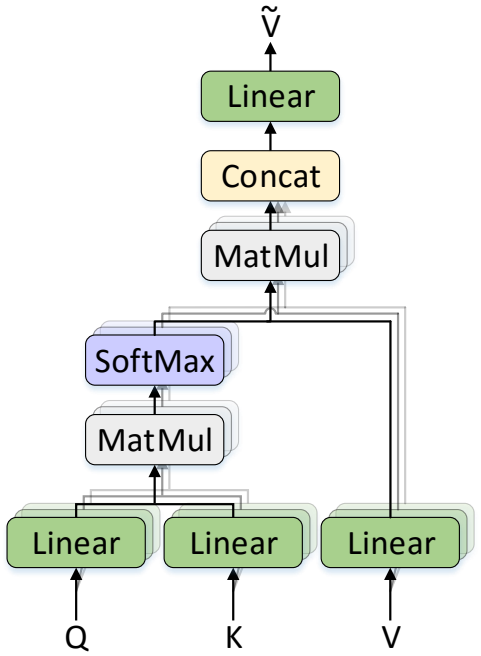
Query (Q): hidden state from language decoder

Keys (K) = Values (V): region-level representations from image encoder

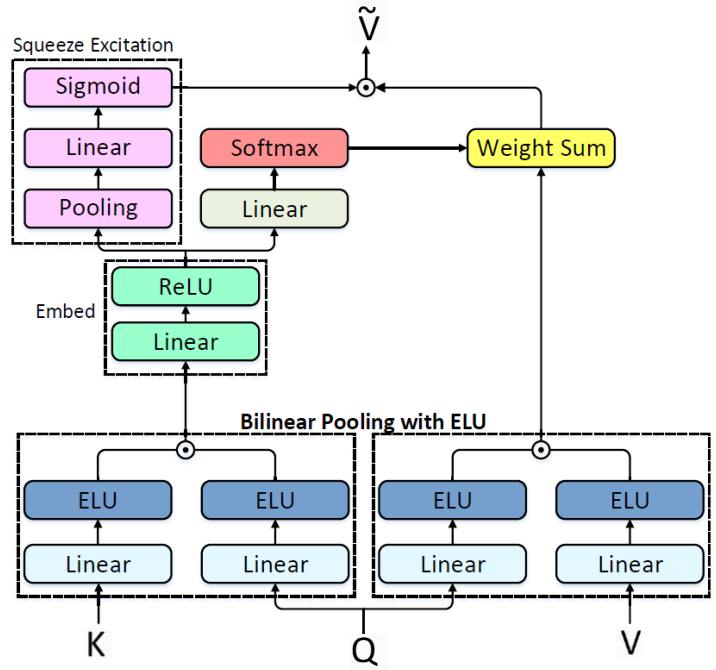
X = visual/top-down attention [Xu, ICML15; Peter, CVPR18]



X = multi-head attention [Sharma, ACL18]



X = x-linear attention [Pan, CVPR20]



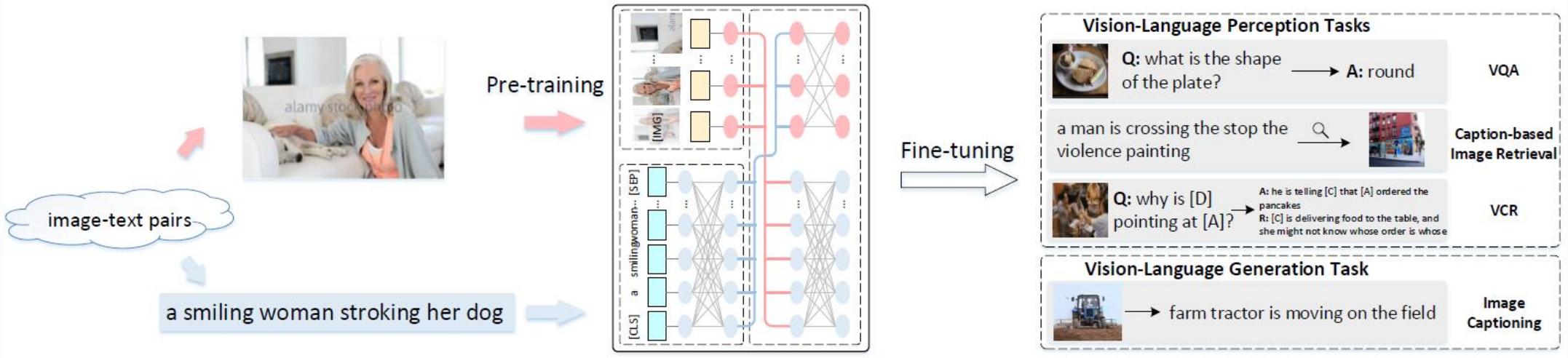
Evaluations on COCO test server [March, 2020]

Model	Group	B@4		METEOR		ROUGE-L		CIDEr-D	
		c5	c40	c5	c40	c5	c40	c5	c40
X-LAN	**Pan, et al., CVPR'20**	**40.3**	**72.4**	**29.6**	**39.2**	**59.5**	**75.0**	**131.1**	**133.5**
HIP	Yao, et al., ICCV'19	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
AoANet	Huang, et al., ICCV'19	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
GCN-LSTM	Yao, et al., ECCV'18	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
RFNet	Jiang, et al., ECCV'18	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
Up-Down	Anderson, et al., CVPR'18	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
LSTM-A	Yao, et al., , ICCV'17	35.6	65.2	27	35.4	56.4	70.5	116	118
Watson Multimodal	Rennie, et al., CVPR'17	34.4	63.6	26.8	35.3	55.9	70.4	112.3	114.6
G-RMI	Liu, et al., ICCV'17	33.1	62.4	25.5	33.9	55.1	69.4	104.2	107.1
MetaMind/VT_GT	Lu, et al., CVPR'17	33.6	63.7	26.4	35.9	55	70.5	104.2	105.9
DLTC@MSR	Gan, et al., CVPR'17	33.1	63.1	25.7	34.8	54.3	69.6	100.3	101.3
reviewnet	Yang, et al., NIPS'16	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9

Code:

Direction III: vision-language pre-training

- Vision-language Pre-training
  - Pre-train multi-modal encoder representation on large-scale vision-language benchmarks
  - Fine-tune multi-modal encoder on vision-language downstream tasks (e.g., VQA, VCR...)

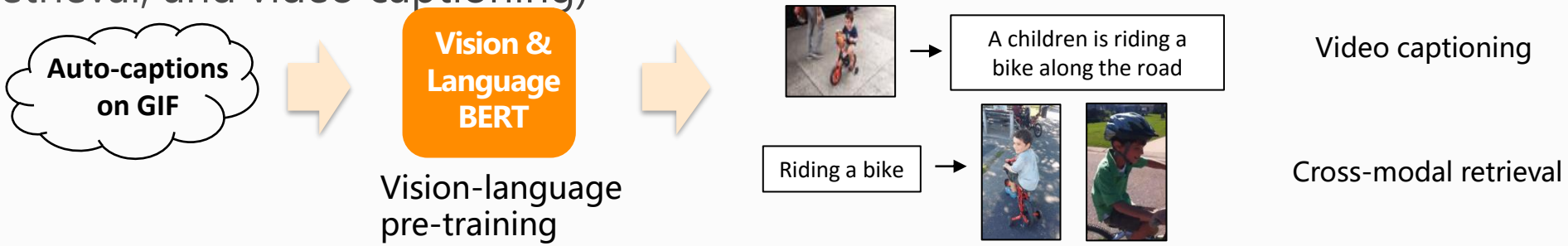


- Dataset
  - Image-Language: Conceptual Captions [Sharma, ACL18, Google]
  - Video-Language: **Auto-captions on GIF** [Pan, Arxiv20, JD AI Research]
- Technology / Open questions
  - How to design a good encoder-decoder structure and proxy tasks?
  - How to pre-train a structure for both VL understanding/generation downstream tasks?

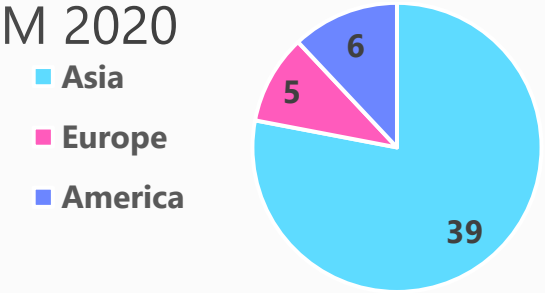
Pre-training for Video Captioning Challenge

http://www.auto-video-captions.top/2020/

- Task Description
  - Leverage **Auto-CapTIONs (ACTION 1.0) on GIF** benchmark to boost research on an emerging task of visual-language pre-training for downstream tasks (e.g., cross-modal retrieval, and video captioning)



- Challenge
  - 50 teams registered challenge
  - 10 teams submitted results
  - Awards will be announced at ACM MM 2020



Rank	Team	Organization	B@4	M	C	S
1	Old Boys	Tsinghua University Beijing University of Posts and Telecommunications Shanghai Ocean University	21.14	17.38	24.42	5.65
2	SYSU-CS	Sun Yat-Sen University	20.41	17.02	23.80	5.39
3	IVIPC-King	University of Electronic Science and Technology of China	18.24	16.46	21.36	5.25

双排扣大衣略带一点中性的帅气，加上legging，曲线美尽显。

黑色风衣是经典必备的基础款，显瘦显气质，绝对不会穿出错，搭配白色衬衫和深色小脚裤，简单的搭配很显气质，知性而利落。

Scenario I: Fashion Collocation in JD.com 京东搭配购

京东智联云

User

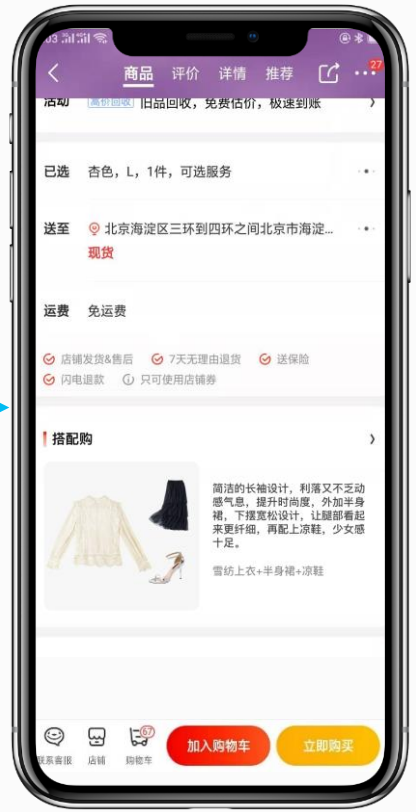
Recommend fashion collocation based on scene and personalized needs

POP Shop

Automatically generate fashion product marketing plan and improve efficiency

Platform

Improve user experience, enhance user browsing time and click



Scenario II: Task-oriented multi-modal dialogue

京东智联云

- Task-oriented dialogue for E-commerce customer service
  - Traditional dialogue: only focus on textual (or voice) modality
  - Multi-modal dialogue: capture semantics across textual and visual (image, video) modalities

这个手机是全面屏吗

是全面屏的呢~小主，您是自己使用呢还是作为礼物送人呢？

自己。我看它的规格那全部写的其他

全面屏手机是目前业界通用的概念，通常是指窄边框和高屏占比的手机。它的正面屏幕占了几乎所有的面积，拥有超窄的边框。

他的性能如何，可以吃鸡么？

小主，麦芒7搭载麒麟710八核处理器，与智慧化EMUI 8.2系统高效结合，手机久用都畅快，视频、游戏体验更佳哦。

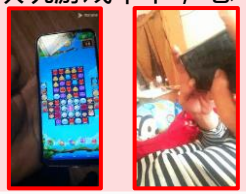
Traditional task-oriented dialogue

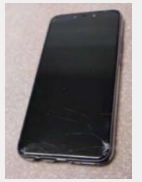
这个手机是全面屏吗

是全面屏的呢~小主，它的正面屏幕占了几乎所有的面积，拥有超窄的边框。视觉体验显著提高：

他的性能如何，可以吃鸡么？

小主，它的性能超级棒的，手机久用都畅快，游戏超级快，很多客户都反映玩游戏不卡，吃鸡很流畅！





小主这是晒的华为麦芒7呢。

客服人员仔细一看小主的此款手机屏幕碎了一个角，您的愁容让我心碎，给我一次机会，为您排忧解难。

建议您在日常的使用中妥善保管好，并且也可以给其套上合适的皮套哦。为了能够更好的协助您处理，将为您联系官方旗舰店客服，并提供合适的修复方案哟。

Task-oriented **multi-modal** dialogue

Language to Vision

Multi-objects Scene

Attention 1.0
- [Mansimov, ICLR'16]
- text-image alignment, low visual quality

Video generation
- TGANs-C [Pan, MM'17]
- low-resolution / short duration / digits motions

Semantic map, Scene-graph
- [Hong, CVPR'18]
- [Johnson, CVPR'18]
- multi-objects scene

Layout generation
- [Hinz, ICLR'19]
- [Song, CVPR'19]
- [Li, CVPR'19]

Relationship
- [Hua, arixv'20]

2014 → 2016 → 2017 → 2018 → 2019 → 2020

Conditioning
- cGAN [Mirza, arXiv'14]
- flagship work / digits / low resolution

Text encoding
- [Reed, ICML'16]
- pioneer work / low resolution

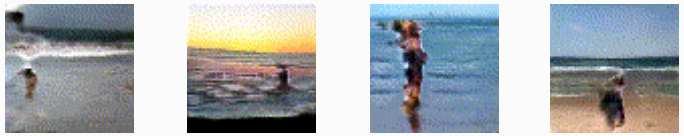
Stacked architecture
- StackGAN [Zhang, ICCV'17]
- StackGAN++[Zhang, arXiv'17]
- high-resolution

Attention 2.0
- AttnGAN [Xu, CVPR'18]
- DA-GAN [Ma, CVPR'18]
- sub-region details

Redescription
- [Qiao, CVPR'19]
- text -> image -> text'

Single Object

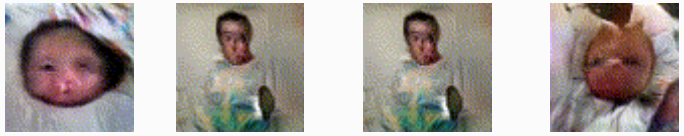
Direction I: Caption to Video



beach

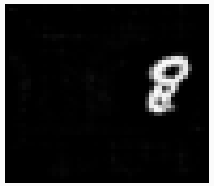
golf

train station



baby

[Vondrick, NIPS'16]



digit 8 is moving left and right

digit 2 is up and down and digit 0 is left and right

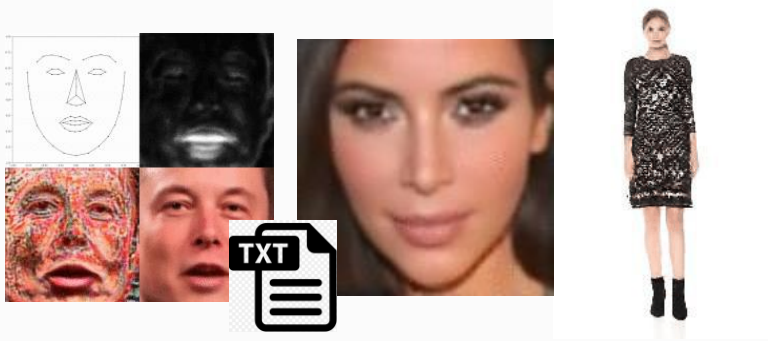
a cook puts noodles into some boiling water

a person is cutting beef

[Pan, Yao, Mei, ACM MM'17]

a long hair man who wears a black vest is playing Guitar

[Yu, arxiv'20]



Talking head/full-body Generation

[Chen, CVPR'19, Wang & Mei, MM'20]

Direction II: Multi-objects Scene

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

a herd of sheep grazing on a lush green field

a fruit stand display with bananas and kiwi

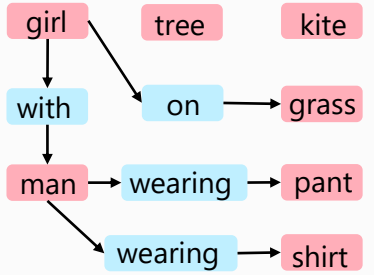
a young girl eating a slice of pizza

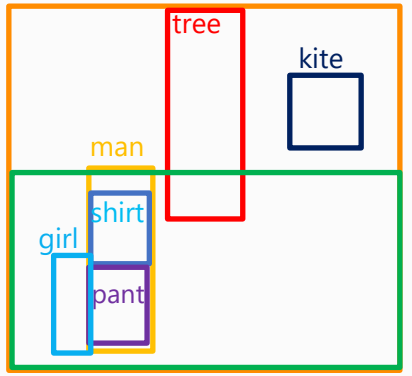
[Xu, Zhang, Huang, Zhang, Gan, Huang, He, CVPR'18]

A woman, man and a dog standing in the snow

Two rectangular slices of pizza with multiple toppings

[Hinz, Heinrich, Wermter, ICLR'19]





[Hua, Bai, Zhang, Mei, arxiv'20]

Trends & Application

- Trend 1：Generative Pre-Training
  - Background：Lang (GPT-3, promising) + Vision (ImageGPT, emerging)
  - Lang-to-vision is naturally *Generative Pre-Training* for high-level V&L tasks
- Trend 2：Knowledge
  - Background：Knowledge has been successful in language domain (dialog, QA)
  - Incorporating knowledge (besides data statistics) in V&L is important for boosting model generalization, explainability

- Applications
  - 文本到图像的检索
  - 小说/文章的自动配图/插画
  - 数字虚拟人 – 从“相似性”到“互补性”转变的V&L
    - 文本驱动的视觉形象（文本→表情、唇形、动作）
    - 用于播报、对话场景，如客服/虚拟主播/带货