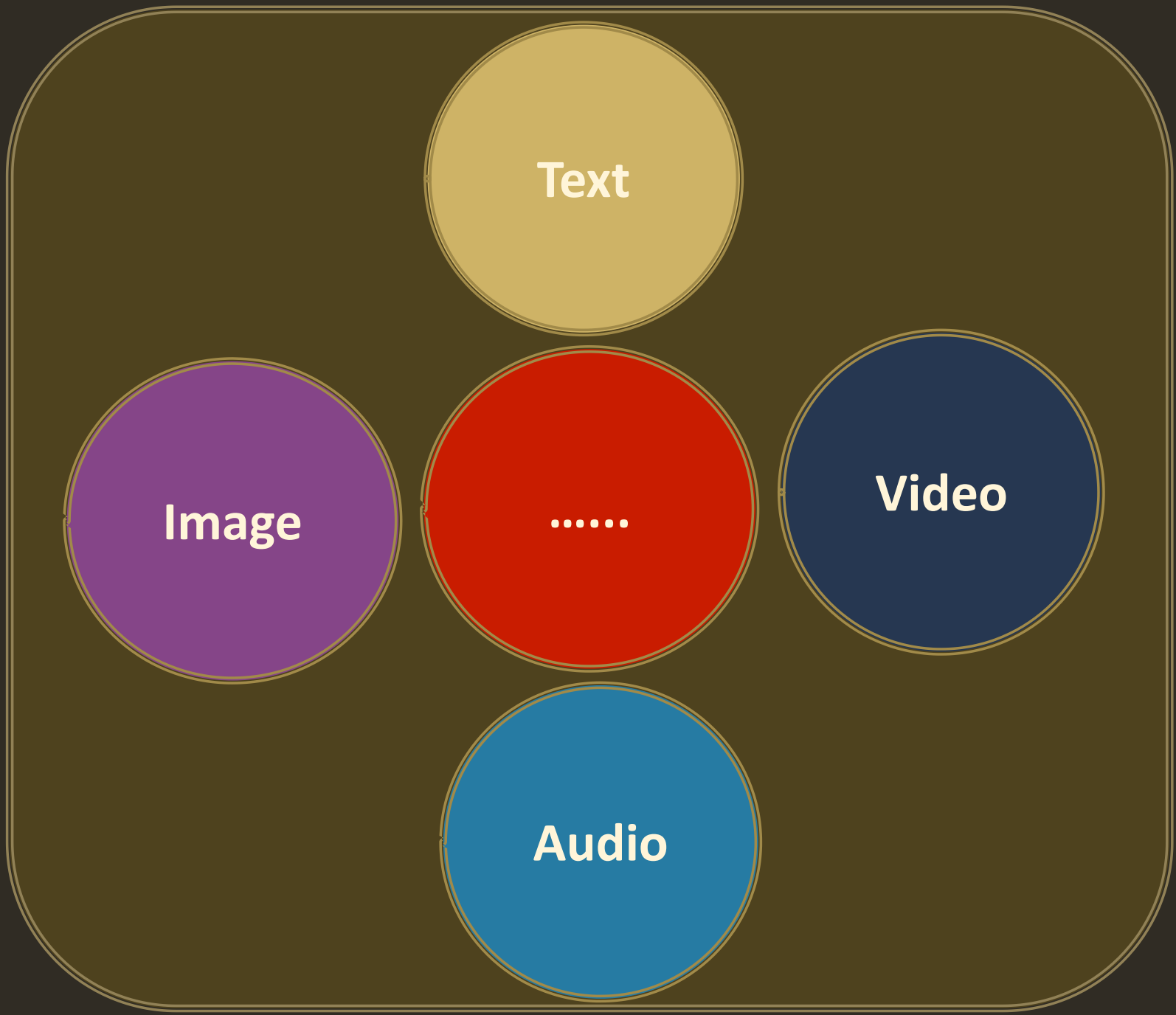
Cross-media Intelligence:

Vision and Language

Prof. Heng Tao Shen （申恒涛）
School of Computer Science and Engineering
AI Research Institute & Center for Future Media
University of Electronic Science and Technology of China

未来媒体研究中心 CENTER FOR FUTURE MEDIA | 电子科技大学 University of Electronic Science and Technology of China

* Let’s start with ...


Text
Image
.....
Video
Audio

The Future is Multimedia

SECURITY

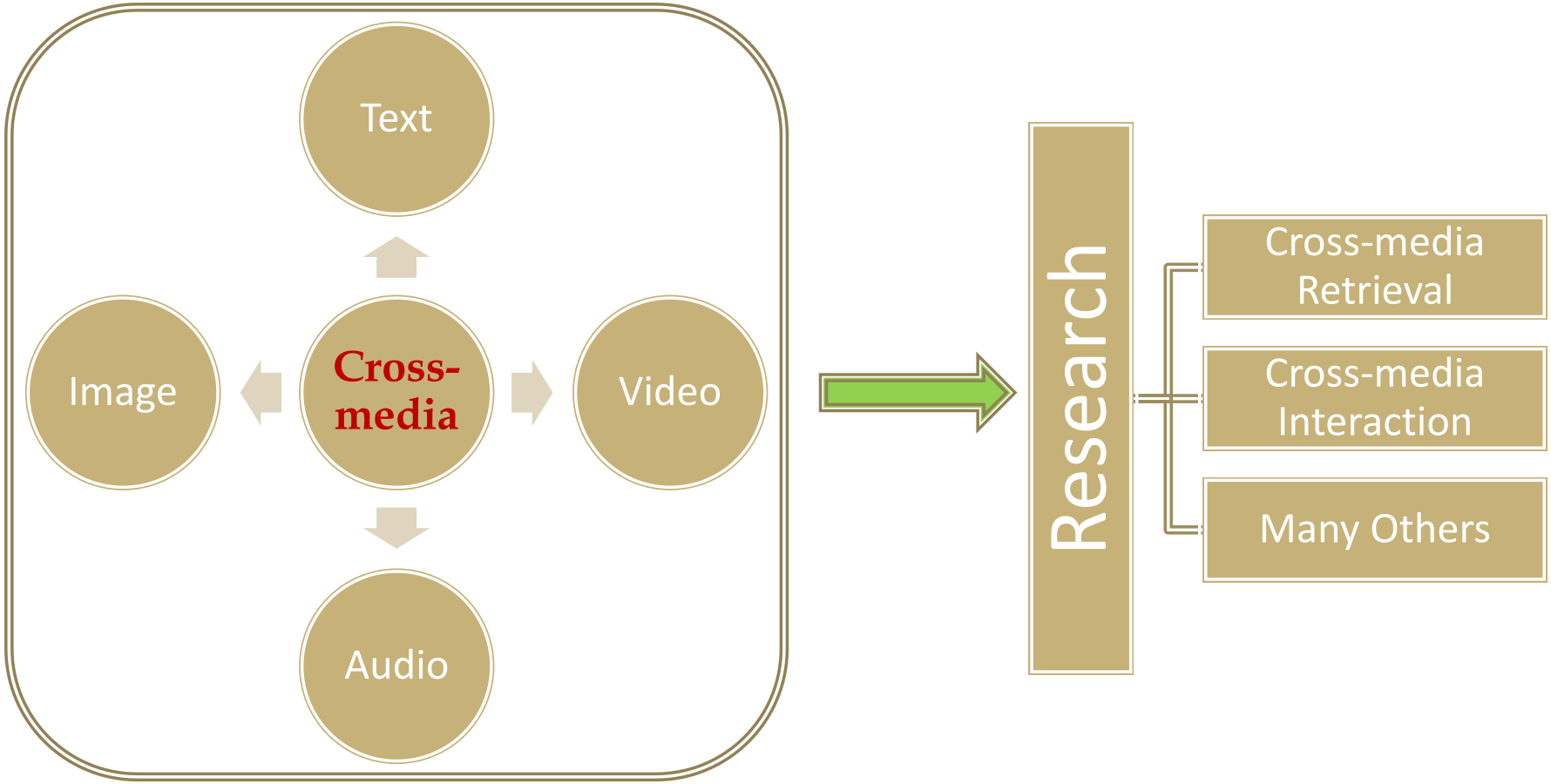
DRONES/AERIAL

AUTOMOTIVE

CONSUMER

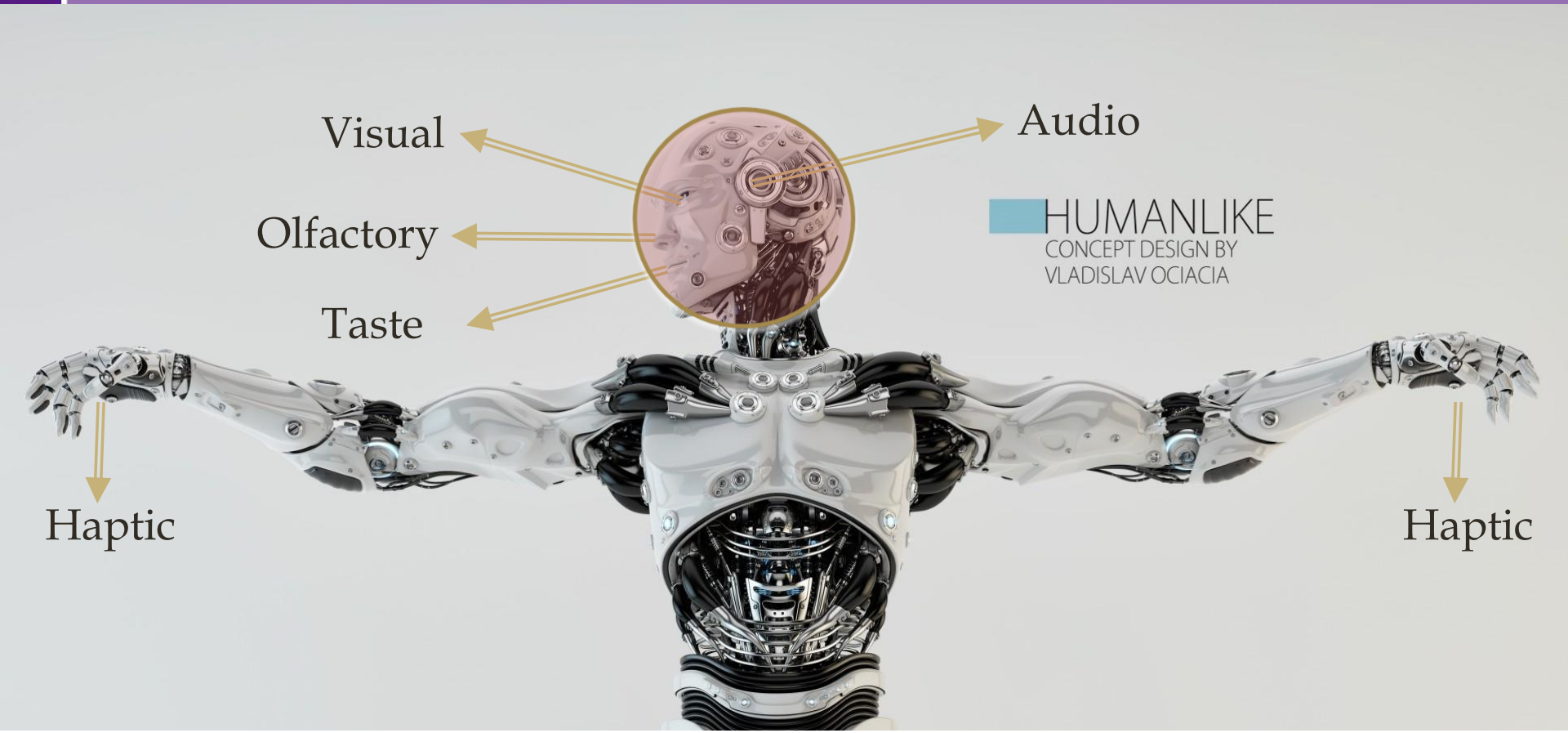
- By 2020 (IHS), security cameras will capture
  - 30 billion images per second
  - 100 trillion images pre hour
  - 2.5 million terabyte per day
  - 29 cameras per person
- Crucial for security, economy and spans all types of applications

Fully utilize multimedia? cross-media research



Exploit the relationships between different media types

An example: robot



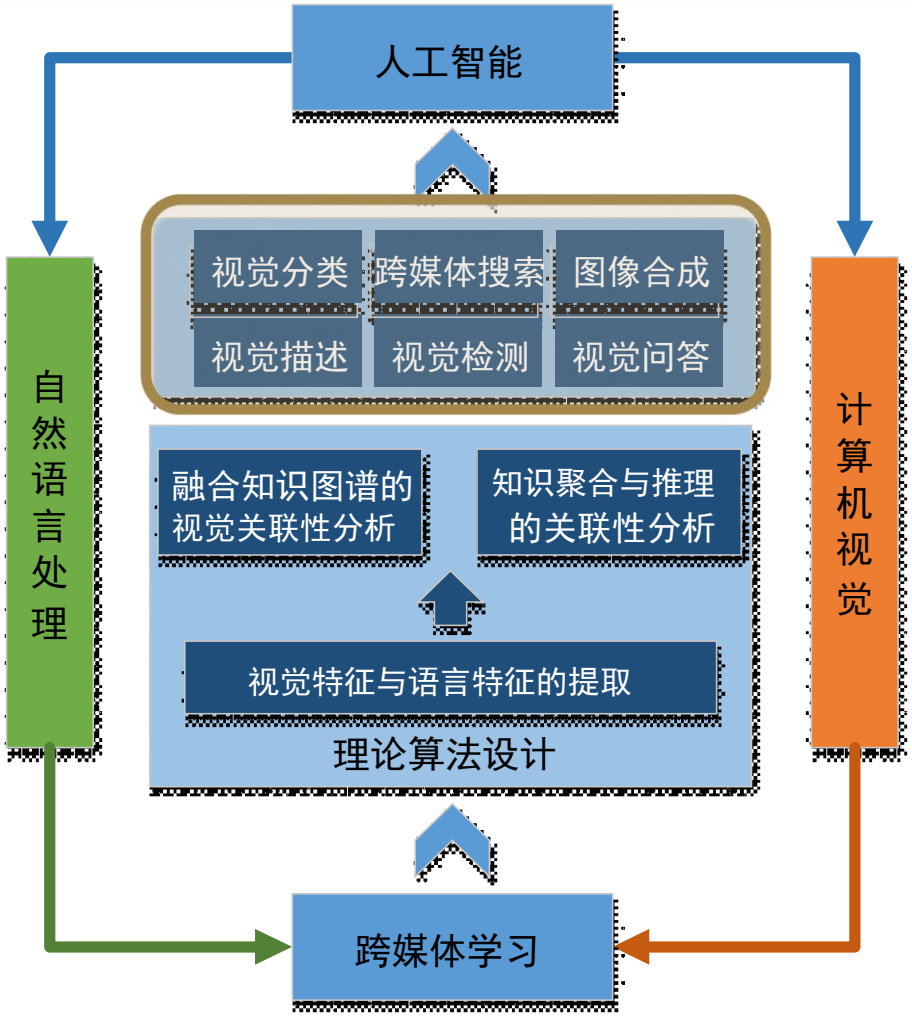
Multimedia data analyzed as a whole for a robot to think/act like human

中国 AI 2.0 - 5大智能方向

- 大数据智能
- 群体智能
- 跨媒体智能
- 混合增强智能
- 自主无人系统

- 推动人工智能发生如下跃变：
  - 从人工知识表达到大数据驱动知识学习
  - 从个体智能到基于互联网络的群体智能
  - **从单一数据到跨媒体认知、学习和推理**
  - 从追求机器智能到人机混合的增强智能
  - 从机器人到自主无人系统的跨越

Vison and Language



Binary classification

Problem Statement

✓ Classify binary data with binary weights

$\mathbf{x} : d \times 1$ → $\mathbf{b} \in \{-1, +1\}^r$

$\mathbf{W} : d \times C$ → $\mathbf{W} \in \{-1, +1\}^{r \times C}$

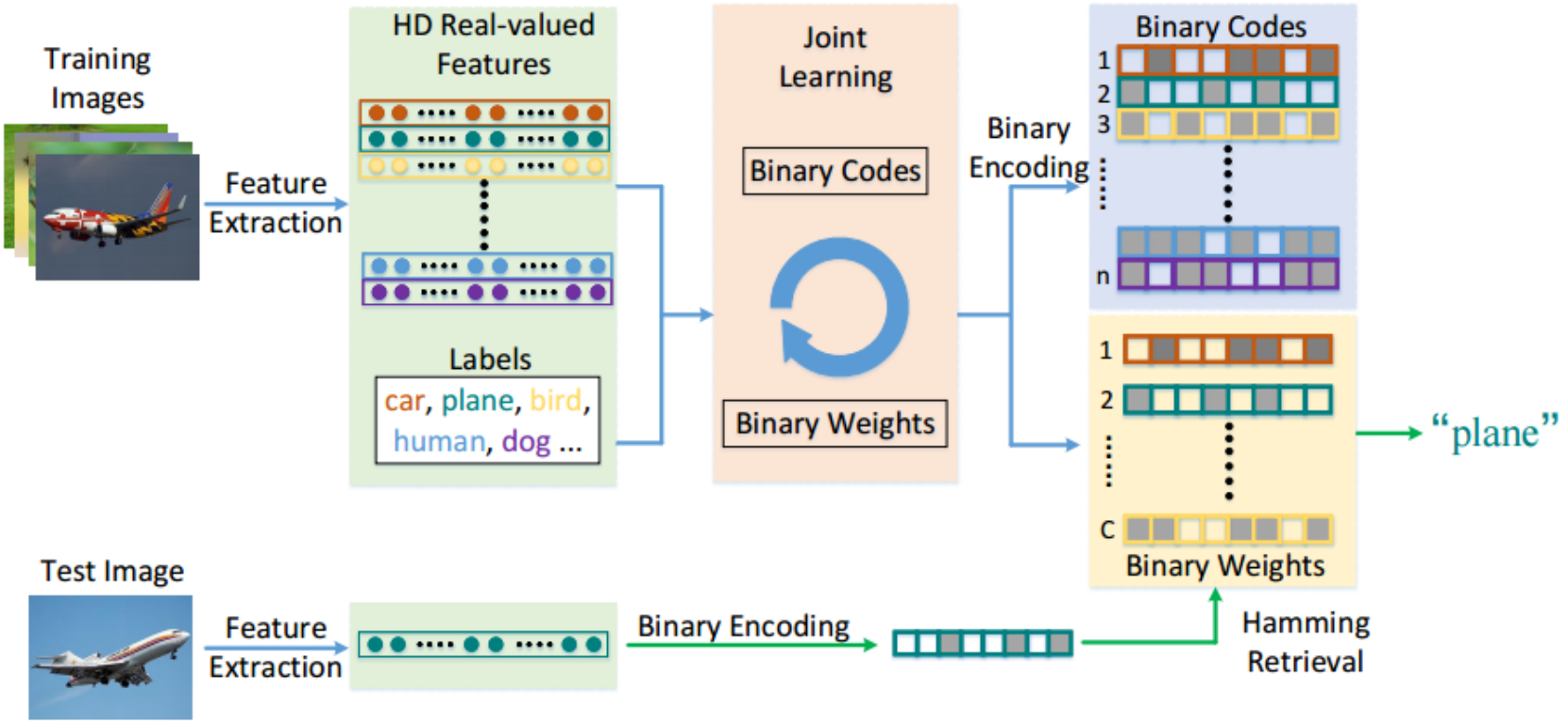
$\mathbf{y} = \mathbf{W}^\top \mathbf{x}$ → $\mathbf{y} = \mathbf{W}^\top \mathbf{b}$

dC Floating-point multiplications → rC XNOR operations

$\mathbf{w}_c^\top \mathbf{b} = r - 2\mathbb{D}_{\mathbb{H}}(\mathbf{w}_c, \mathbf{b})$

$\mathbb{D}_{\mathbb{H}}$: Hamming distance

Binary classification



Classifying an image reduces to retrieving its nearest class codes in the Hamming space

Association for Computing Machinery

SPECIAL INTEREST GROUP ON INFORMATION RETRIEVAL

2017 Annual Meeting

August 7 - 11, 2017 Tokyo, Japan

presents the

BEST PAPER AWARD HONORABLE MENTION

to

Fumin Shen, Yadong Mu, Yang Yang, Wei Liu, Li Liu, Jingkuan Song, Heng Tao Shen

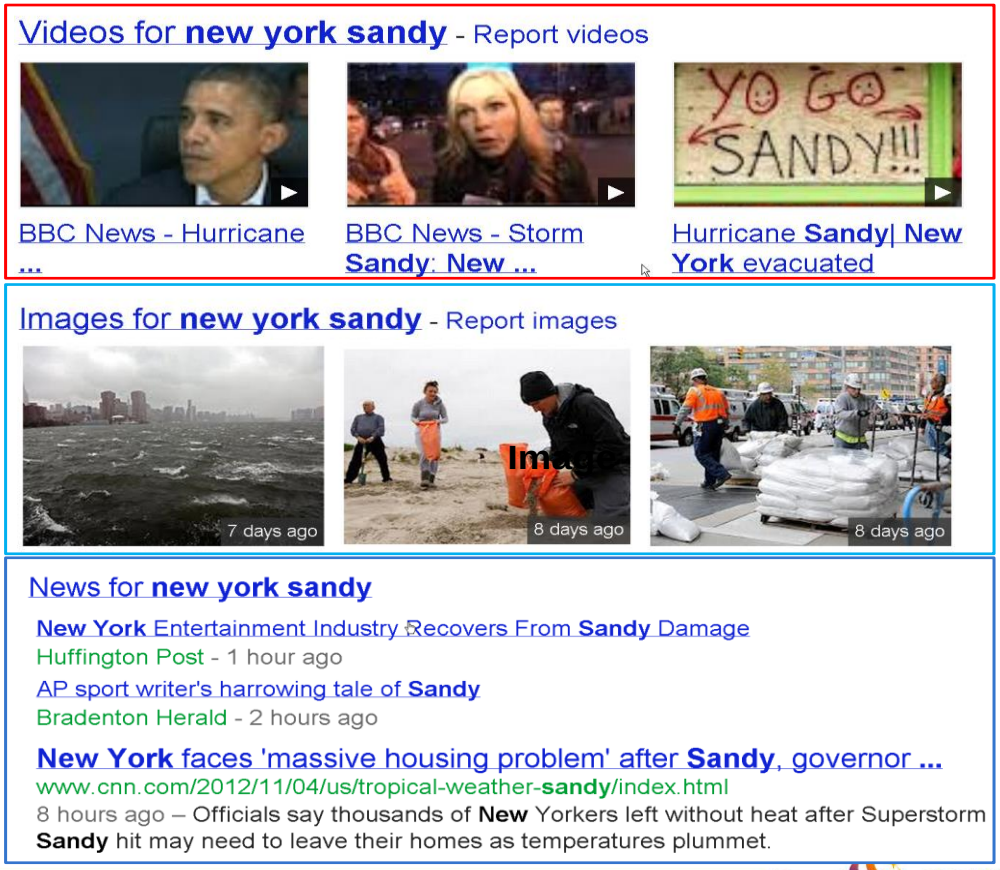
for

Classification by Retrieval: Binarizing Data and Classifiers

Diane Kelly	Noriko Kando	Tetsuya Sakai	Hideo Joho
Chair, SIGIR	Co-Chair, SIGIR 2017	Co-Chair, SIGIR 2017	Co-Chair, SIGIR 2017

Multimedia Data – Multimodal In Nature

- Explosion of multimedia contents that are represented by **multiple modalities** coming from **multiple sources**



Multimodal Data - Heterogeneous in Nature

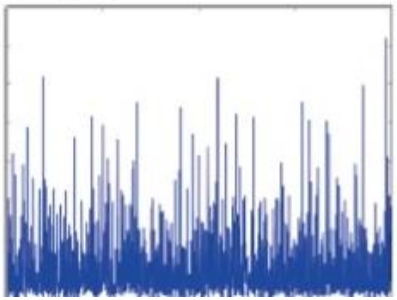
- Images and texts have very different natures of representation

Image

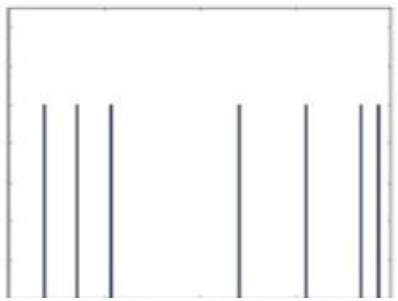
Text

yellow, frog, amphibian, canon, 550d, eos, picture

Dense **real-valued**

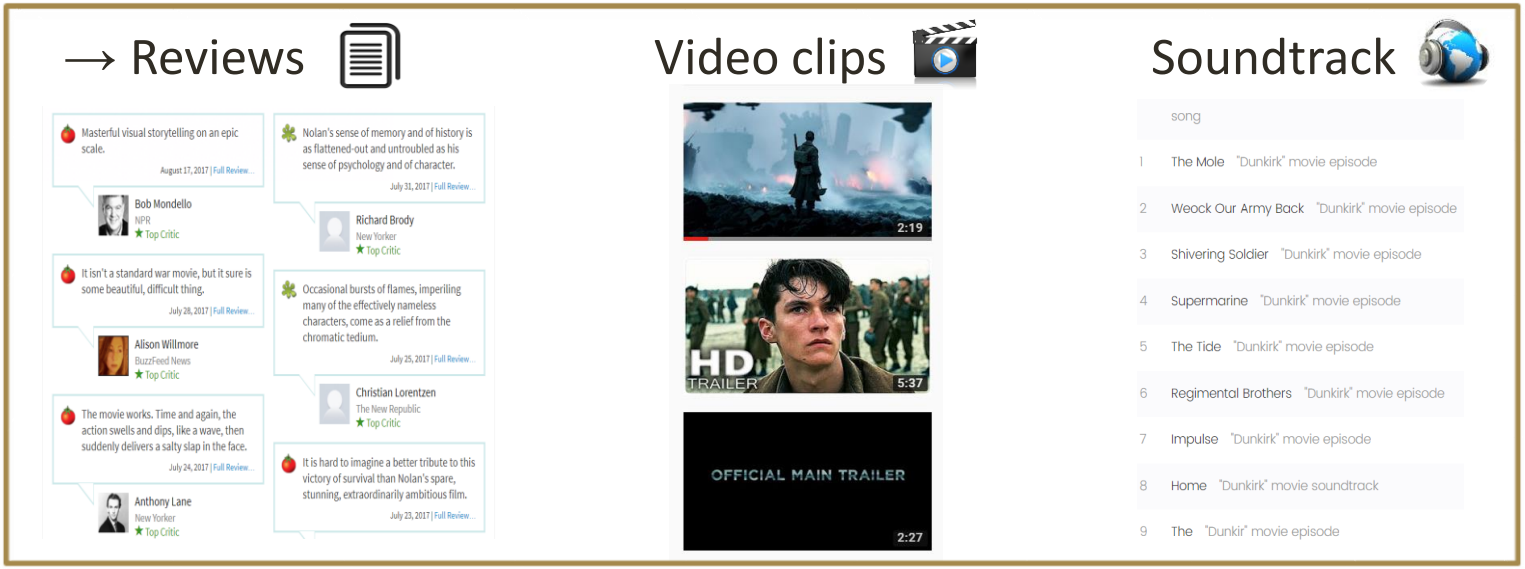


Sparse **discrete**



Cross-Modal Retrieval

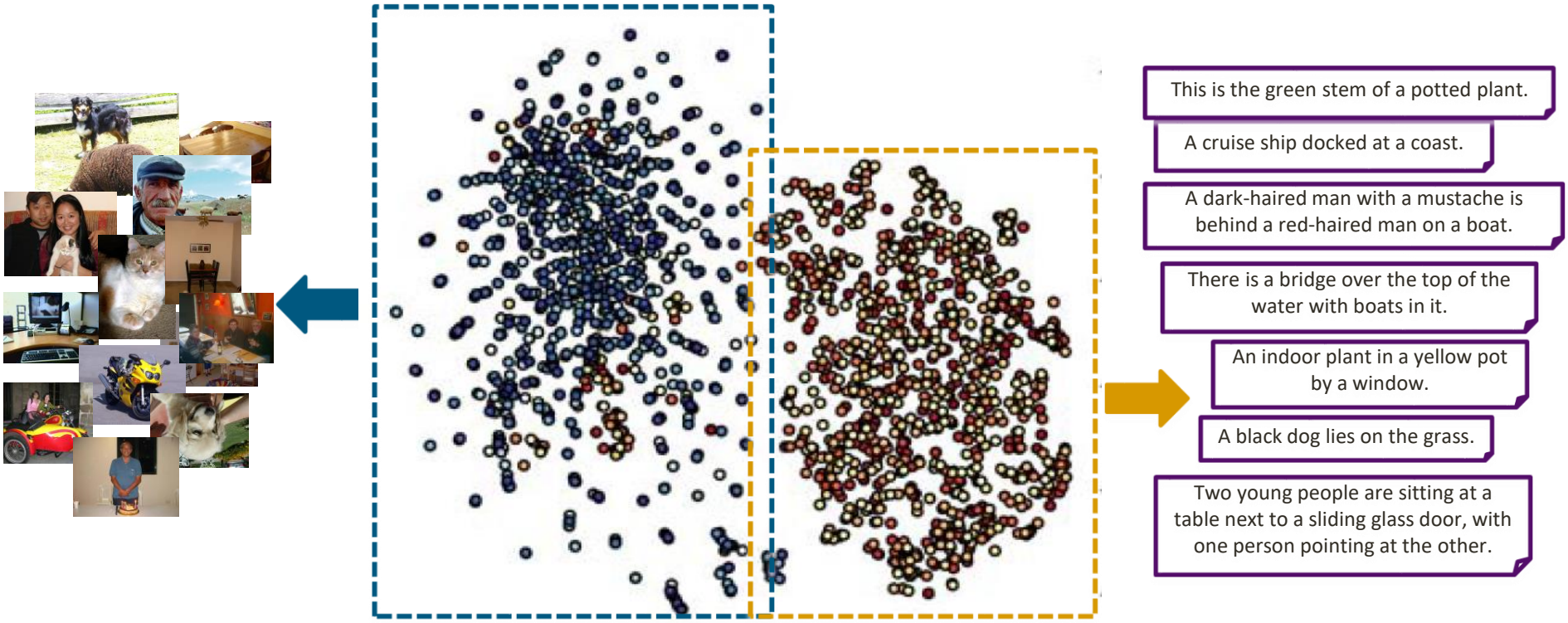
- Strong need for "modality invariant" retrieval system
  - Given the "title" of one movie, e.g. "**Dunkirk**"



- Cross-Modal Retrieval:
  Given a query from one modality, search the closest matches in another modality

Main Challenge

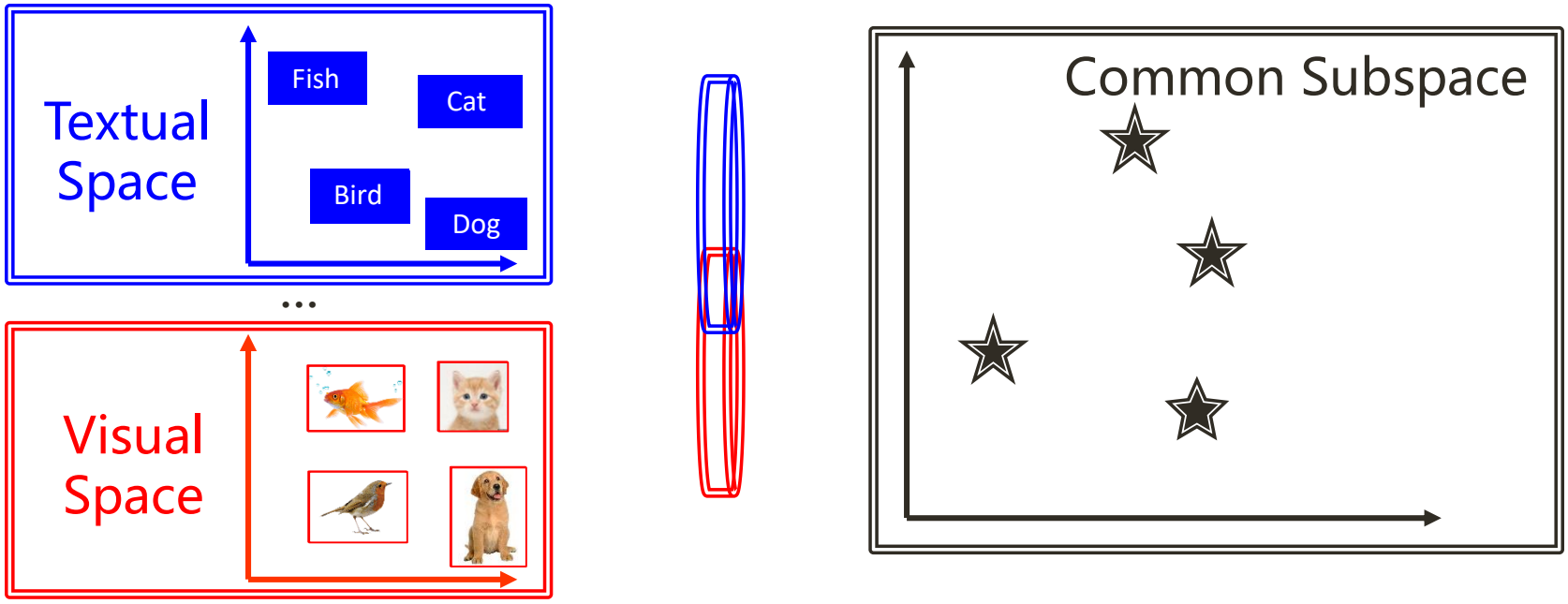
- Heterogeneous representations -> **Modality Gap**
  - Similarities cannot be directly measured
  - Distributions cannot be well aligned



On Flicker-30K dataset

Main Approach: Common Subspace Learning

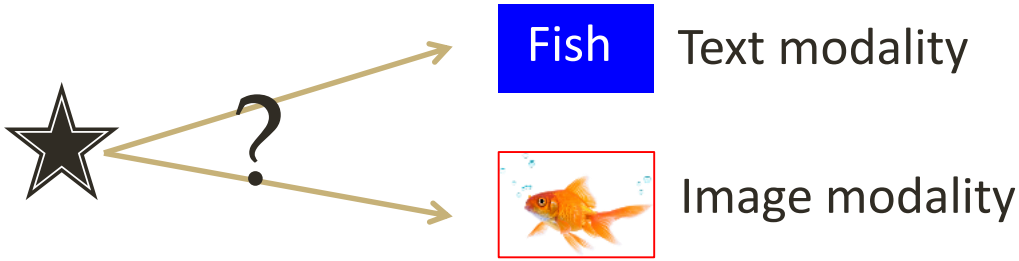
- Main Tasks of Subspace Learning
  - Learning discriminative representations for each modality
  - Modeling inter-item correlations across modalities



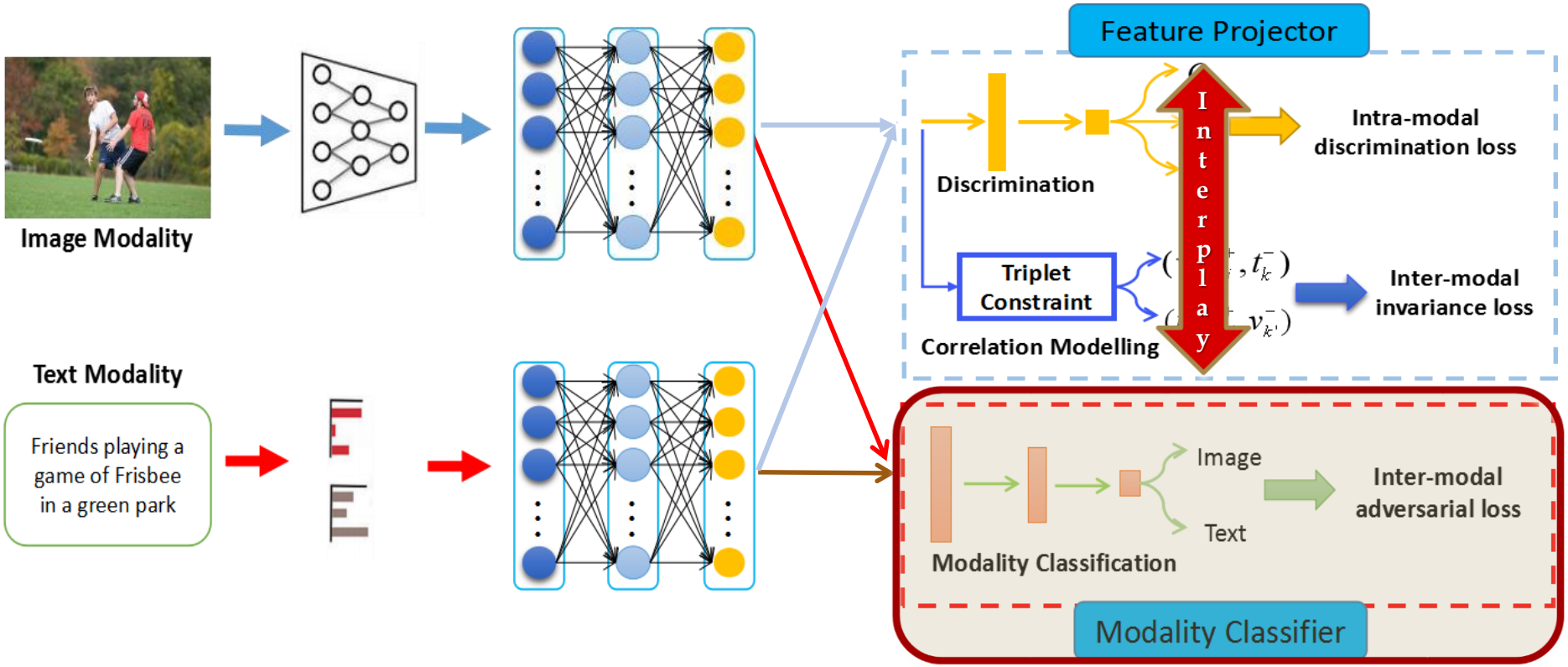
Limitations of Previous Deep Methods

- Limitations
  - Previous methods mainly focus on feature discrimination and pairwise item correlations
  - Modality invariance is seldom considered

  However, it is non-trivial since it is hard to correlate the cross-modal items if the shift between modalities is large
- Ideal case: modality invariant representation
  - Given a subspace representation, its original modality cannot be identified

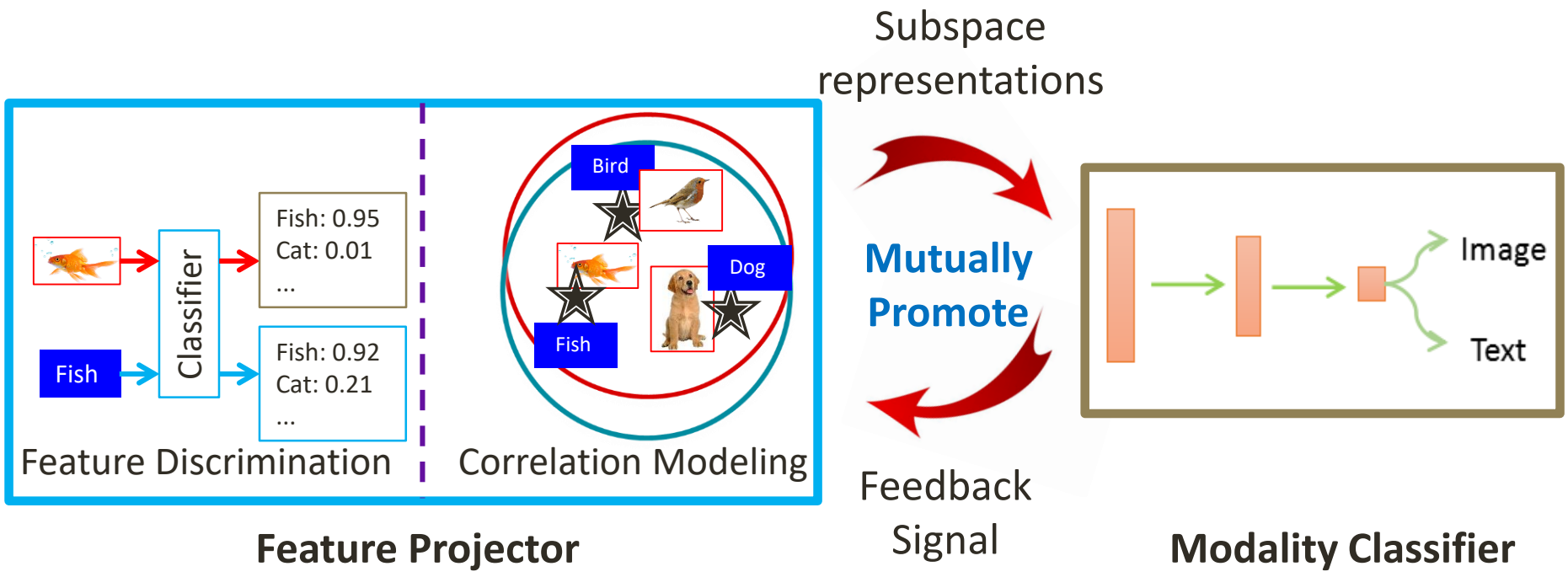


Ours: Adversarial Cross-Modal Retrieval (ACMR)



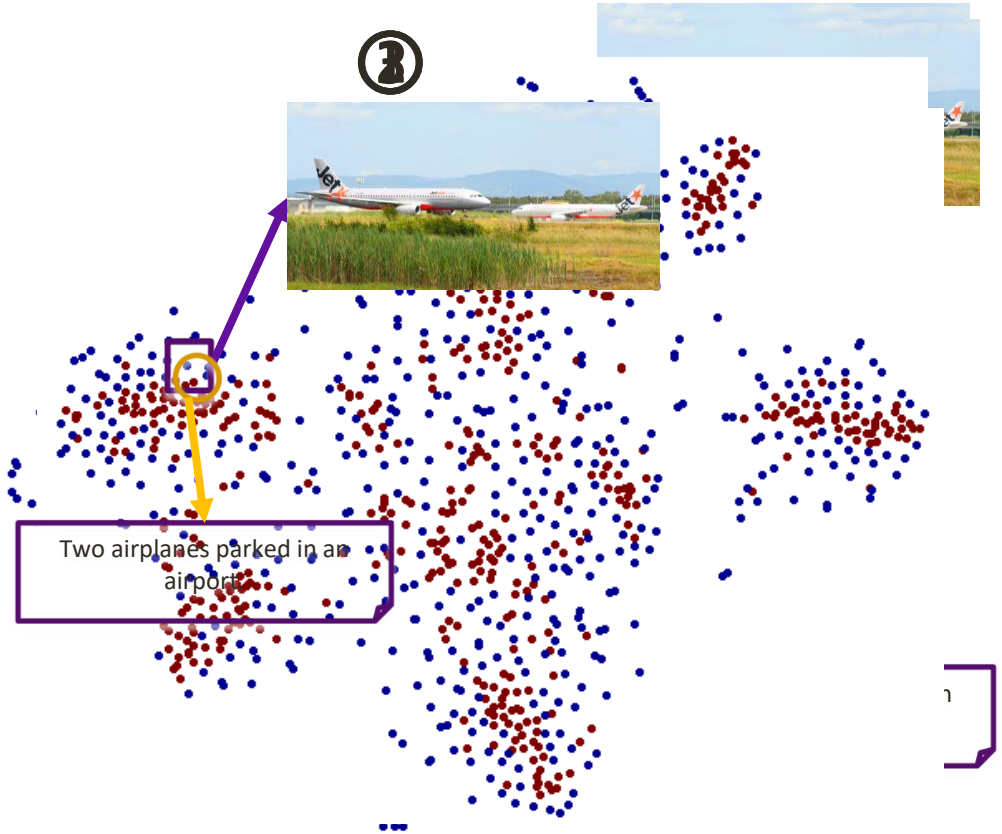
☐ **Adversarial Framework: Interplay** between **feature projector** and **modality classifier**

Novelties of ACMR

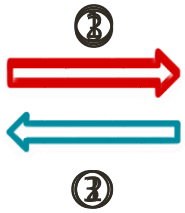


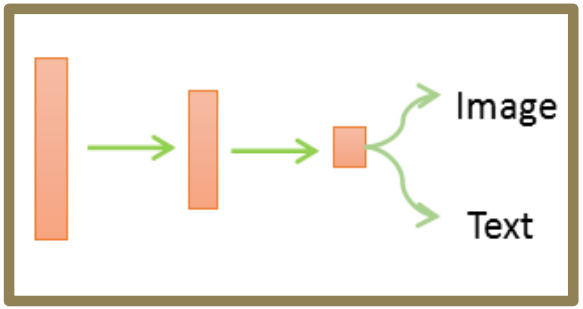
- Introducing **immediate feedback** signal to steer the learning process of feature projector, for mitigating the modality gap
- Simultaneously exploit feature discrimination and correlation modeling

An Example



Feature Projector





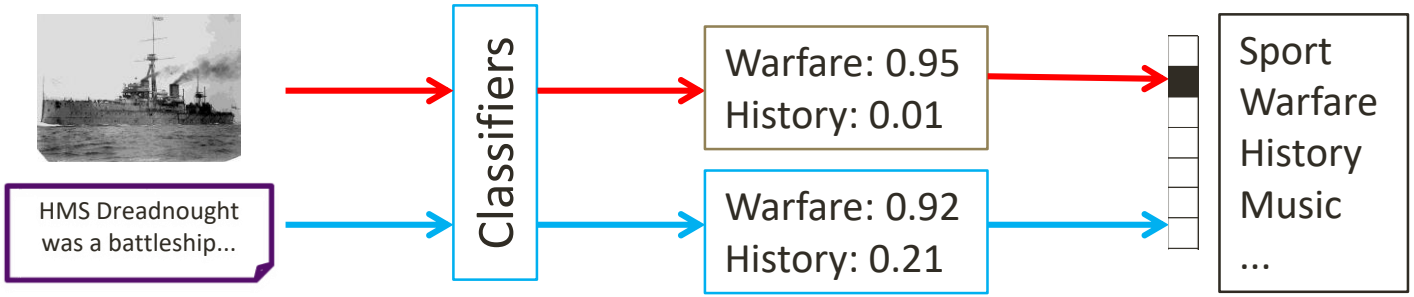
Modality Classifier

Feature Discrimination

- Improved discrimination of subspace representations
  - Image and text classifiers to output predicted probability distributions
  - Supervised by the same set of semantic abstractions

$$\hat{p}_i(v_i, y_i) = \frac{e^{\phi_{y_i}(v_i)}}{\sum_{i=1}^{C} \phi_{y_i}(v_i)} \qquad \hat{p}_i(t_i, y_i) = \frac{e^{\phi_{y_i}(t_i)}}{\sum_{i=1}^{C} \phi_{y_i}(t_i)}$$

$$L_{imd} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot (\log \hat{p}_i(v_i) + \log \hat{p}_i(t_i)))$$

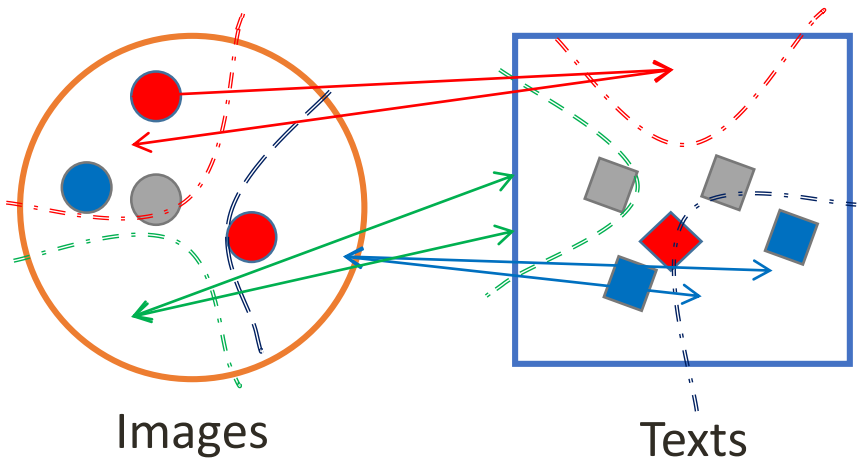


Correlation Modeling

- Goals of Correlation Modeling
  - *Minimize* the distances among *semantically similar* items
  - *Maximize* the distances among *semantically different* items
- Enforce triplet constraints $\{(v_i, t_j^+, t_j^-)\}_i, \{(t_i, v_j^+, v_j^-)\}$

$$L_{imi,V}(\theta_V) = \sum_{i,j,k} (\max(0, \mu + \lambda \cdot l(v_i, t_j^+) - l(v_i, t_j^-)))$$

$$L_{imi,T}(\theta_T) = \sum_{i,j,k} (\max(0, \mu + \lambda \cdot l(t_i, v_j^+) - l(t_i, v_j^-)))$$



Overall Formulation for Feature Projector

$$L_{emb} = \alpha \cdot L_{imi} + \beta \cdot L_{imd} + L_{reg}$$

L_{imi}

L_{imd}

L_{reg}

$L_{imi,V}(\theta_V) + L_{imi,T}(\theta_T)$

Inter-modal correlation loss combining visual and textual modalities

$$-\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot (\log \hat{p}_i(v_i) + \log \hat{p}_i(t_i)))$$

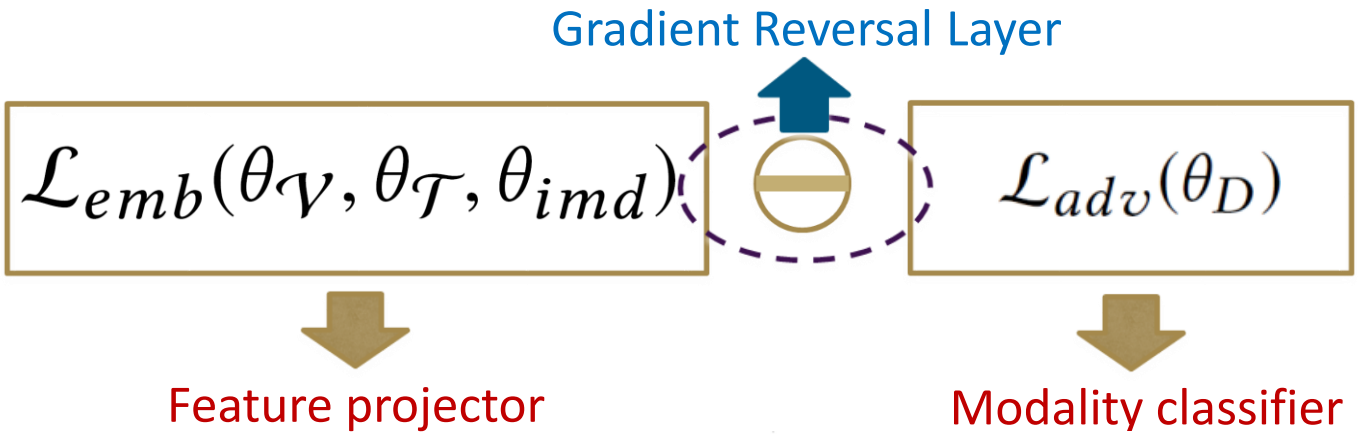
Intra-modal discrimination loss

$$\sum_{l=1}^{L}(||W_v^l||_F + ||W_t^l||_F)$$

Regularization for preventing overfitting

ACMR Optimization: Minimax Game

- Integrate losses from the two "players"



$$\mathcal{L}_{adv}(\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}(\mathbf{m_i} \cdot (\log D(v_i; \theta_D) + \log(1 - D(t_i; \theta_D))).$$

- A minimax game between the two "players"

$$(\hat{\theta}_{\mathcal{V}}, \hat{\theta}_{\mathcal{T}}, \hat{\theta}_{imd}) = \arg\min_{\theta_{\mathcal{V}}, \theta_{\mathcal{T}}, \theta_{imd}} (\mathcal{L}_{emb}(\theta_{\mathcal{V}}, \theta_{\mathcal{T}}, \theta_{imd}) - \mathcal{L}_{adv}(\hat{\theta}_D))$$

$$\hat{\theta}_D = \arg\max_{\theta_D} (\mathcal{L}_{emb}(\hat{\theta}_{\mathcal{V}}, \hat{\theta}_{\mathcal{T}}, \hat{\theta}_{imd}) - \mathcal{L}_{adv}(\theta_D))$$

Experimental Settings

Data and Features

Dataset	Instances	Labels	Image feature	Text feature
Wikipedia	1,300/1,566	10	128d SIFT 4,096d VGG	10d LDA 1,000d BoW
Pascal Sentences	800/200	20	4,096d VGG	1000d BoW
NUSWIDE-10k	8000/2000	350	4,096d VGG	3000d BoW
MSCOCO	66,226/16,557	500	4,096 VGG	3,000 BoW

Configuration, Tasks and Metrics

- Network: V → 2000 → 200 for visual feature; T → 500 → 200 for textual feature; f → 50 → 1 for modality classifier
- Tasks: Img2Txt and Txt2Img
- Metric: Mean Average Precision (MAP) and precision-scope curve

Compared Methods

- Traditional: SCM (CCA) [MM'10], JRL [TCSVT'14], LCFS [ICCV'15], CCA-3V [IJCV'14], JFSSL [TPAMI'16]
- DNN-based: Multimodal-DBN [ICML'12], Bimodal-AE [ICML'11], Corr-AE [MM'14], CMDN [IJCAI'16]

Experimental Results

- Retrieval results on Wikipedia Dataset

	Shallow feature			Deep feature		
	Img2Txt	Txt2Img	Average	Img2Txt	Txt2Img	Average
CCA	0.255	0.185	0.220	0.267	0.222	0.245
Mitimodal DBN	0.149	0.150	0.150	0.204	0.183	0.194
Bimodal-AE	0.236	0.208	0.222	0.314	0.290	0.302
CCA-3V	0.275	0.224	0.249	0.437	0383	0.410
LCFS	0.279	0.214	0.246	0.455	0.398	0.427
Corr-AE	0.280	0.242	0.261	0.402	0.395	0.398
JRL	0.344	0.277	0.311	0.453	0.400	0.426
JFSSL	0.306	0.228	0.267	0.428	0.396	0.412
CMDN	-	-	-	0.488	0.427	0.458
ACMR	**0.366**	**0.277**	**0.322**	**0.619**	**0.489**	**0.546**

✓ Adversary for modality invariance modeling $\Leftrightarrow L_{adv}$

Experimental Results

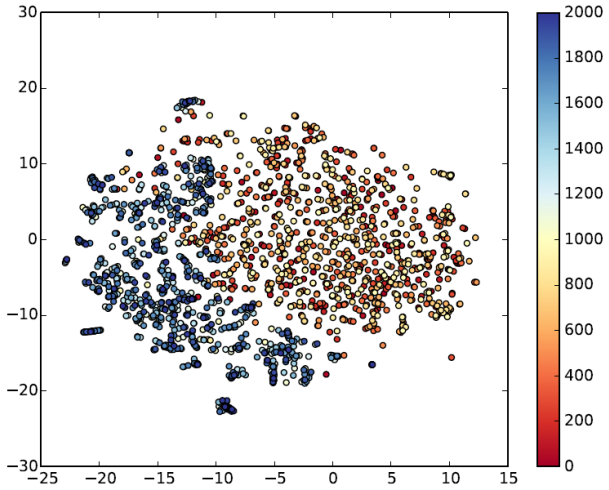
- Retrieval results on Pascal Sentences and NUSWIDE-10k

	Pascal Sentences			NUSWIDE-10k		
	Img2Txt	Txt2Img	Average	Img2Txt	Txt2Img	Average
CCA	0.363	0.219	0.291	0.189	0.188	0.189
Multimodal DBN	0.477	0.424	0.451	0.201	0.259	0.230
Bimodal AE	0.456	0.470	0.458	0.327	0.369	0.348
LCFS	0.442	0.357	0.400	0.383	0.346	0.365
Corr-AE	0.489	0.444	0.467	0.366	0.417	0.392
JRL	0.504	0.489	0.496	0.426	0.376	0.401
CMDN	0.534	0.534	0.534	0.492	0.515	0.504
ACMR	**0.535**	**0.543**	**0.539**	**0.544**	**0.538**	**0.541**

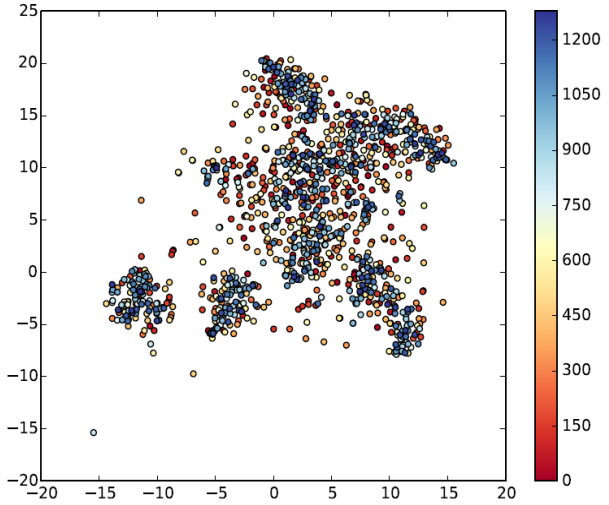
Our approach consistently achieves better performance

Visualization of Subspace Representations

- t-SNE visualizations of subspace representations on Wikipedia



(a) Without Adversary



(b) With Adversary

Adversary is effective to reduce the modality gap

Let's Make History!

Oct 23, 2017 – Oct 27, 2017 • Mountain View, CA USA

ACM Special Interest Group on Multimedia

presents to

Bokun Wang, Yang Yang, Xing Xu,
Alan Hanjalic, Heng Tao Shen

Best Paper Award 2017

"Adversarial Cross-Modal Retrieval"

October, 2017

Video captioning

- Problem Statement

  Describing video visual content with natural language text

A Video: 

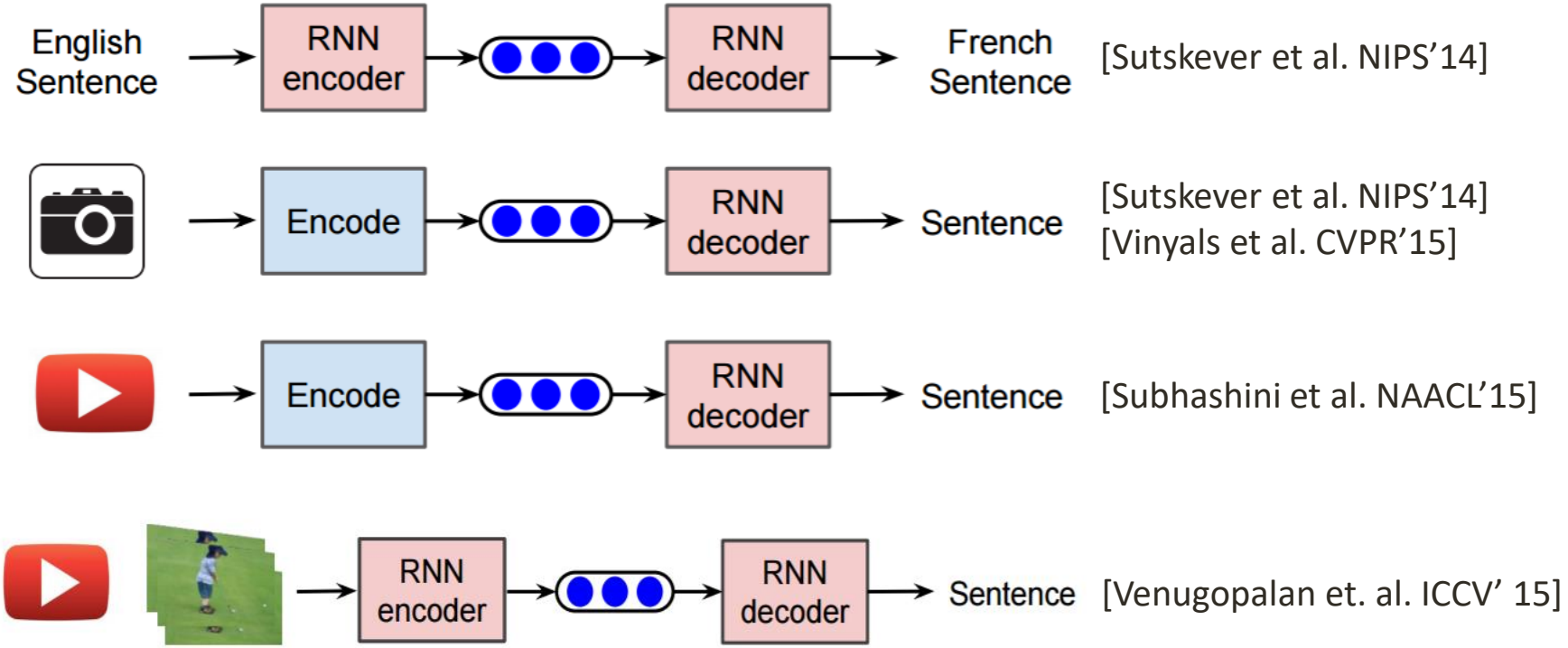
Caption: A man is running down a road

A Video: 

Caption: A monkey is pulling a dog's tail and is chased by the dog.

Related work

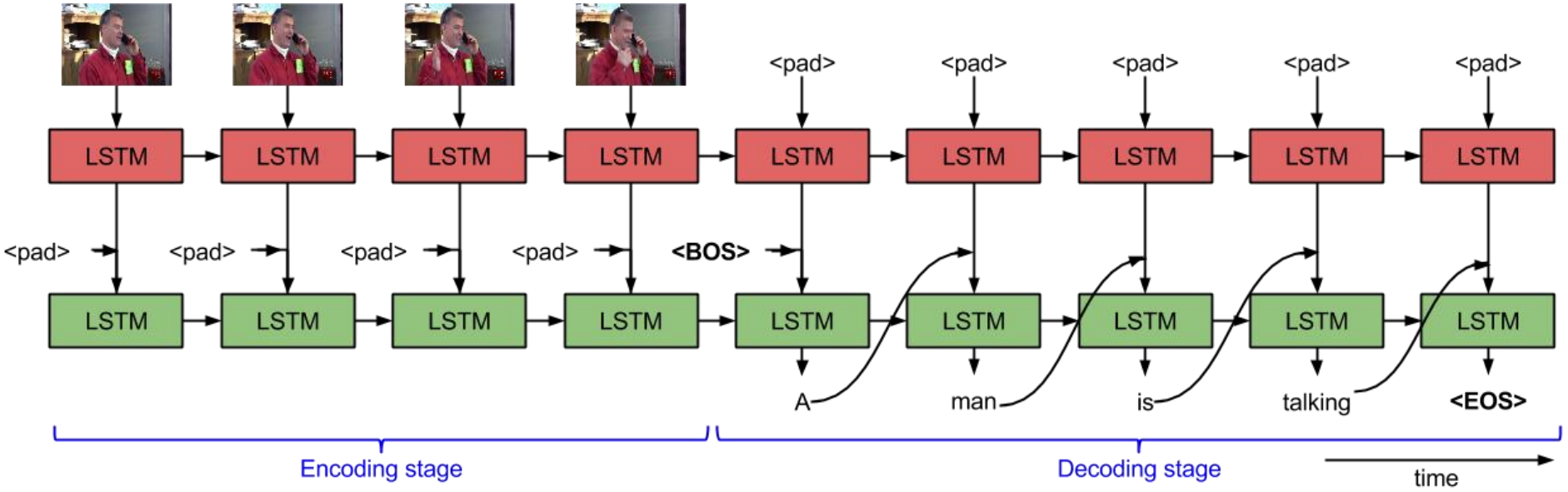
The Encoder-Decoder framework



Key: Encode the visual feature of the video and "decode" it to a sentence

Related work

A basic model for Video Captioning: Sequence to Sequence - Video to Text [1]

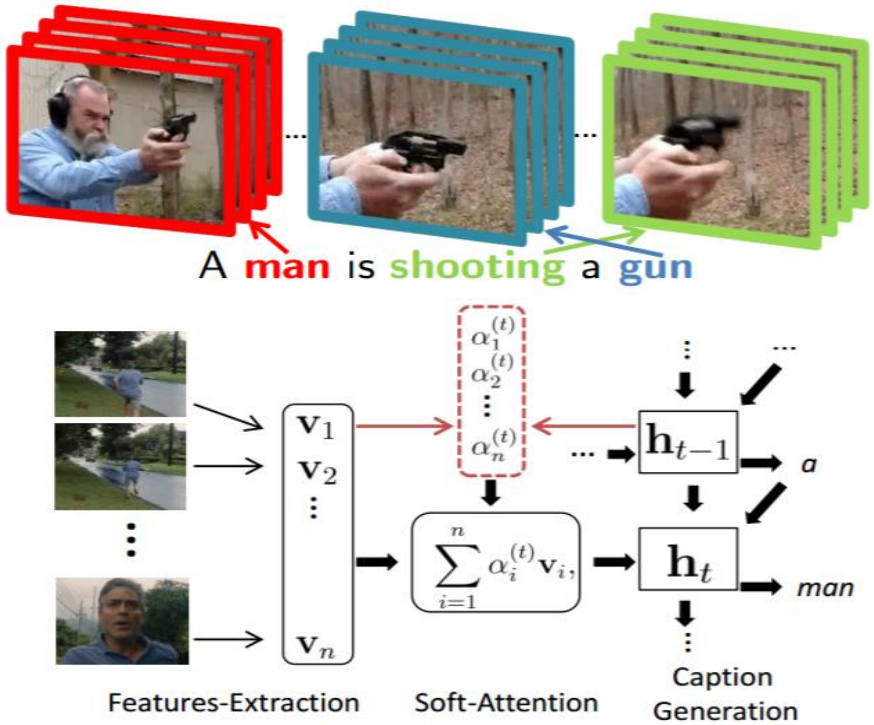


Based on many to many **Hierarchical** RNN： seq of frames -> seq of words

[1] Venugopalan, Subhashini, et al. "Sequence to sequence-video to text.” ICCV 2015.

Related work

Attention Mechanism for Video Captioning [1]



Based on many to many RNN with attention： seq of frames -> seq of words

[1] Yao, Li, et al. "Describing videos by exploiting temporal structure." ICCV 2015.

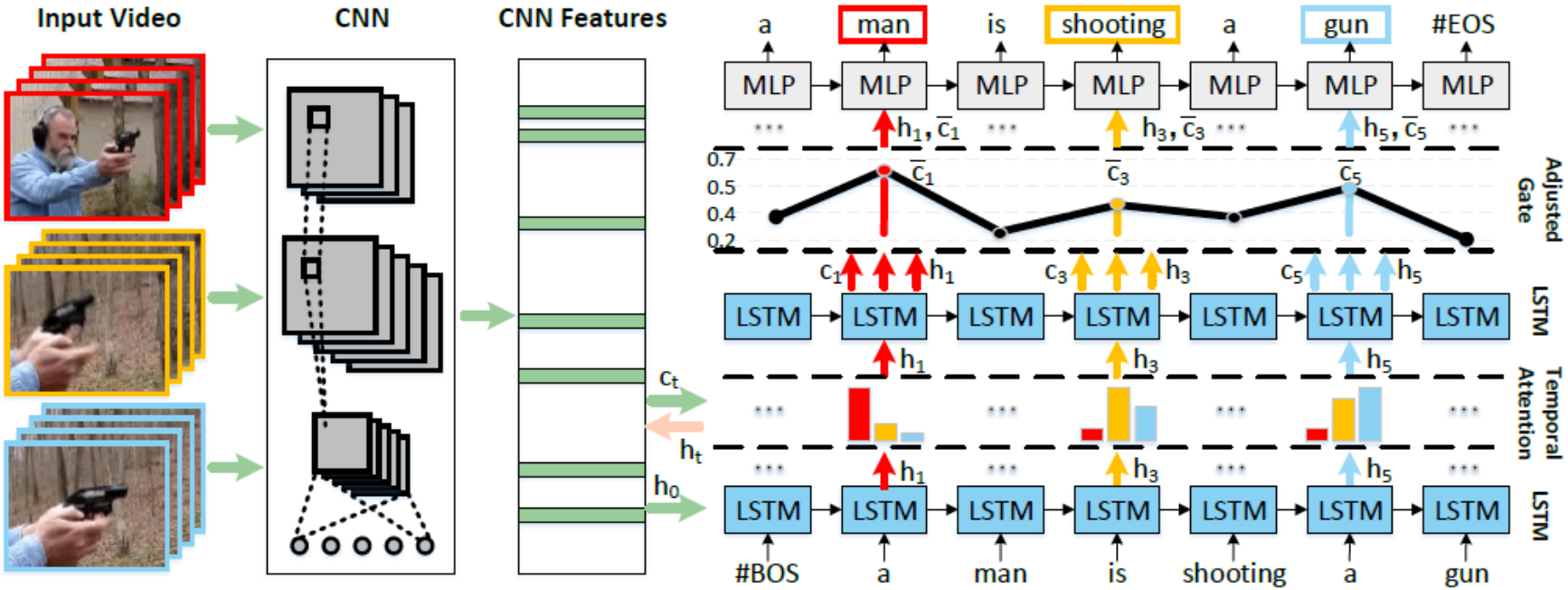
hLSTM with adjusted temporal attention (IJCAI 2017 & TPAMI 2019)

A Video:

Video Caption: a **man** is **shooting** a **gun** #EOS

- **Motivations**
  - Most existing decoders apply the attention mechanism to **every** generated word including non-visual words (e.g. “the”, “a”).
  - However, these non-visual words can be easily predicted using natural language model. Imposing attention mechanism on non-visual words could mislead and decrease the overall performance of video captioning.

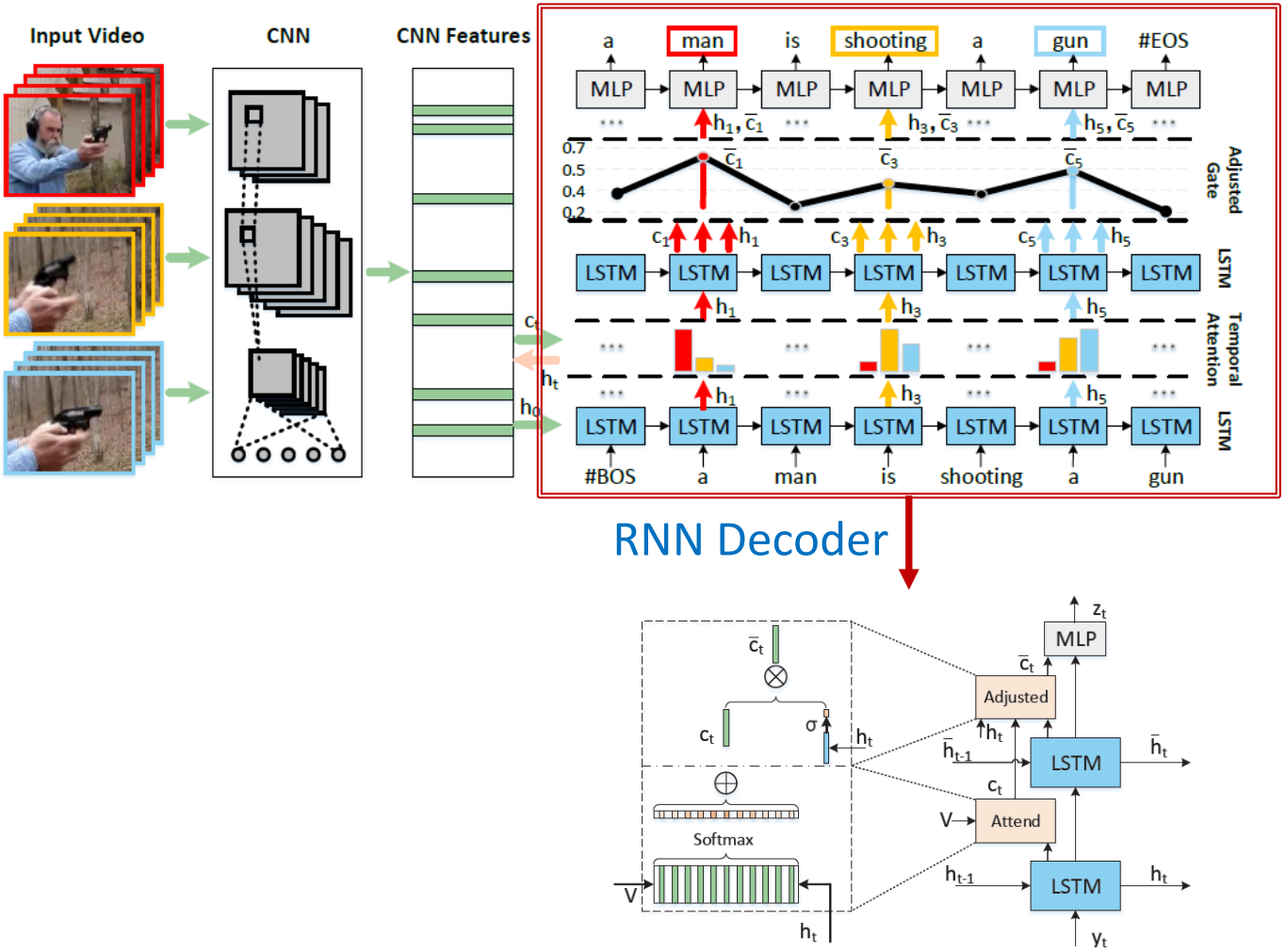
hLSTM with adjusted temporal attention



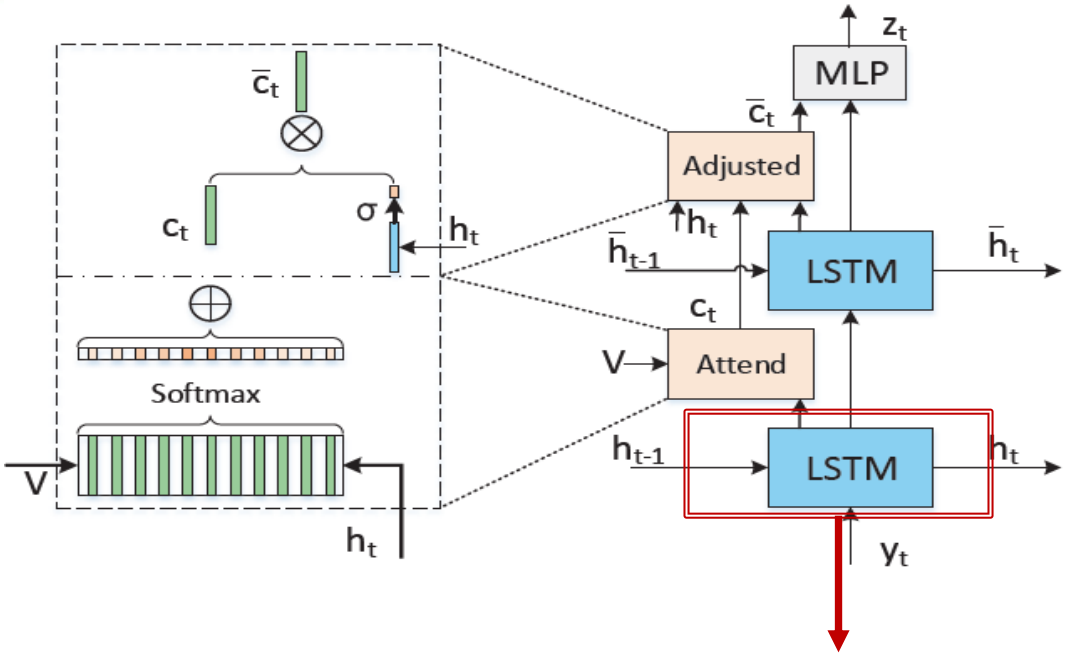
A unified encoder-decoder framework:

- Visual words (e.g. “shooting” or “gun”) are generated with visual information extracting from a set of specific frames.
- non-visual words (e.g. “a” and “is”) are relying on the language model.

hLSTM with adjusted temporal attention



hLSTM with adjusted temporal attention



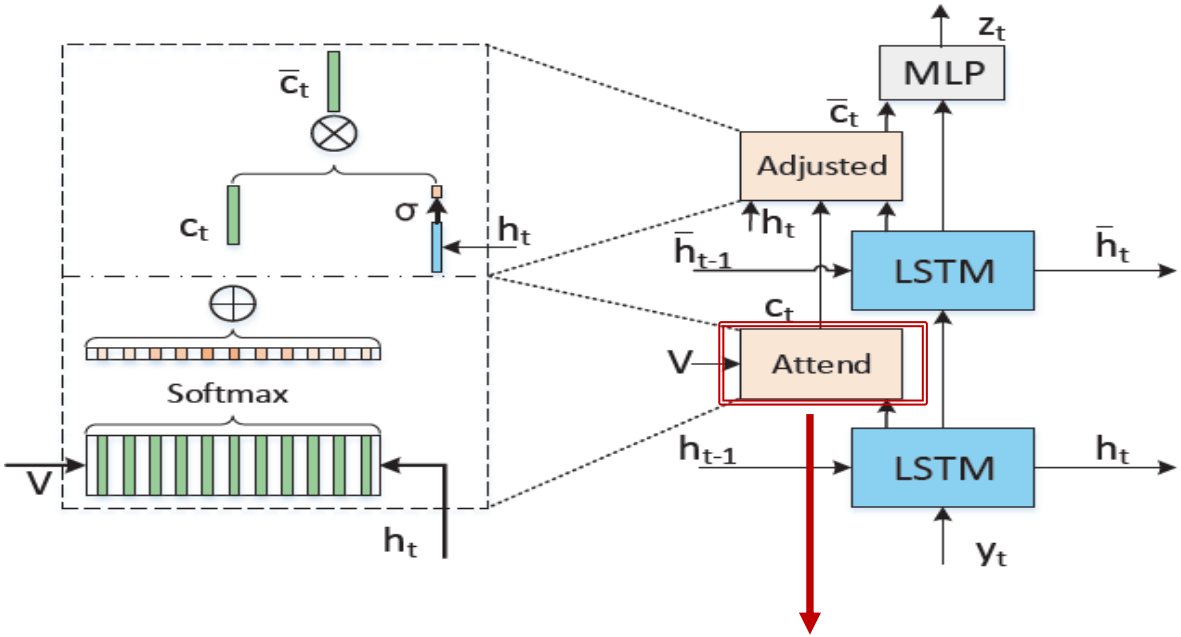
Bottom LSTM

captures the low-level visual features

- Bottom LSTM Layer. For the bottom LSTM layer, the updated internal hidden state depends on the current word y_t, previous hidden state $\mathbf{h}_{t-1}$ and memory state $\mathbf{m}_{t-1}$:

$$\mathbf{h}_0, \mathbf{m}_0 = \left[\mathbf{W}^{\mathrm{ih}}; \mathbf{W}^{\mathrm{ic}}\right] Mean(\{\mathbf{v_i}\})$$
$$\mathbf{h}_t, \mathbf{m}_t = LSTM(\mathbf{y}_t, \mathbf{h}_{t-1}, \mathbf{m}_{t-1})$$

hLSTM with adjusted temporal attention



Temporal Attention Model

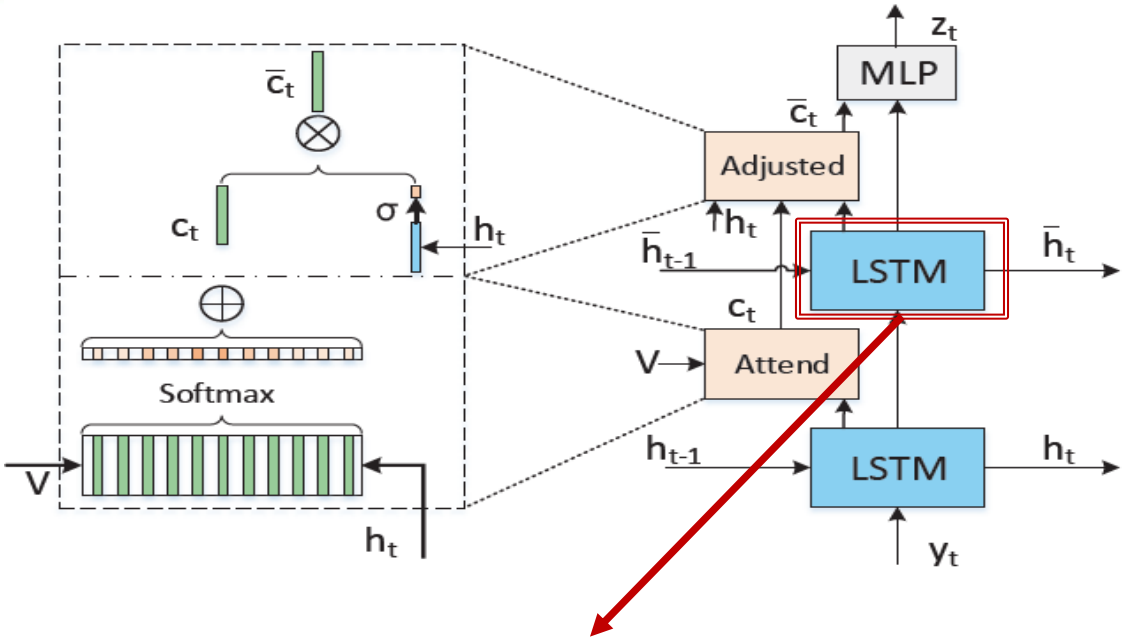
captures the attended visual feature

$$\mathbf{c}_t = \frac{1}{n}\sum_{i=1}^{n} \alpha_t^i \mathbf{v}_i$$

$$\varepsilon_t = \mathbf{w}^T tanh\left(\mathbf{W}_a \mathbf{h}_t + \mathbf{U}_a \mathbf{V} + \mathbf{b}_a\right)$$

$$\alpha_t = softmax(\varepsilon_t)$$

hLSTM with adjusted temporal attention



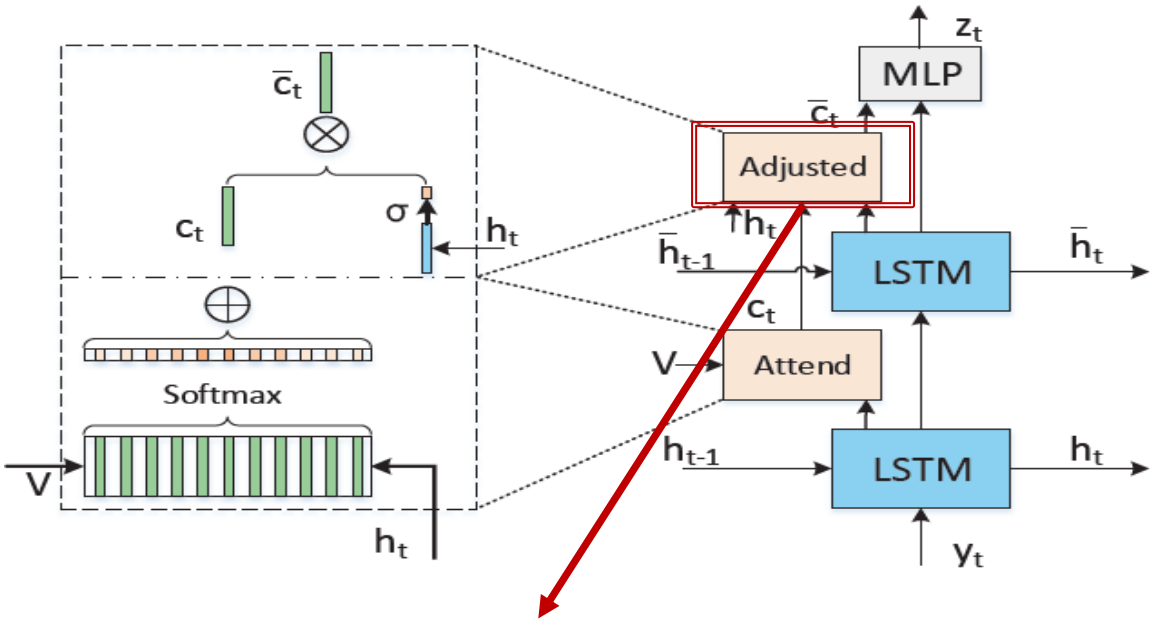
Top LSTM

captures the high-level semantic features of the videos

- Top LSTM Layer. For the top LSTM, it takes the output of the bottom LSTM unit output $\mathbf{h}_t$, previous hidden state $\bar{\mathbf{h}}_{t-1}$ and the memory state $\bar{\mathbf{m}}_{t-1}$ as inputs to obtain the hidden state $\bar{\mathbf{h}}_t$ at time t, and it can be defined as below:

$$\bar{\mathbf{h}}_t, \bar{\mathbf{m}}_t = LSTM(\mathbf{h}_t, \bar{\mathbf{h}}_{t-1}, \bar{\mathbf{m}}_{t-1})$$

hLSTM with adjusted temporal attention

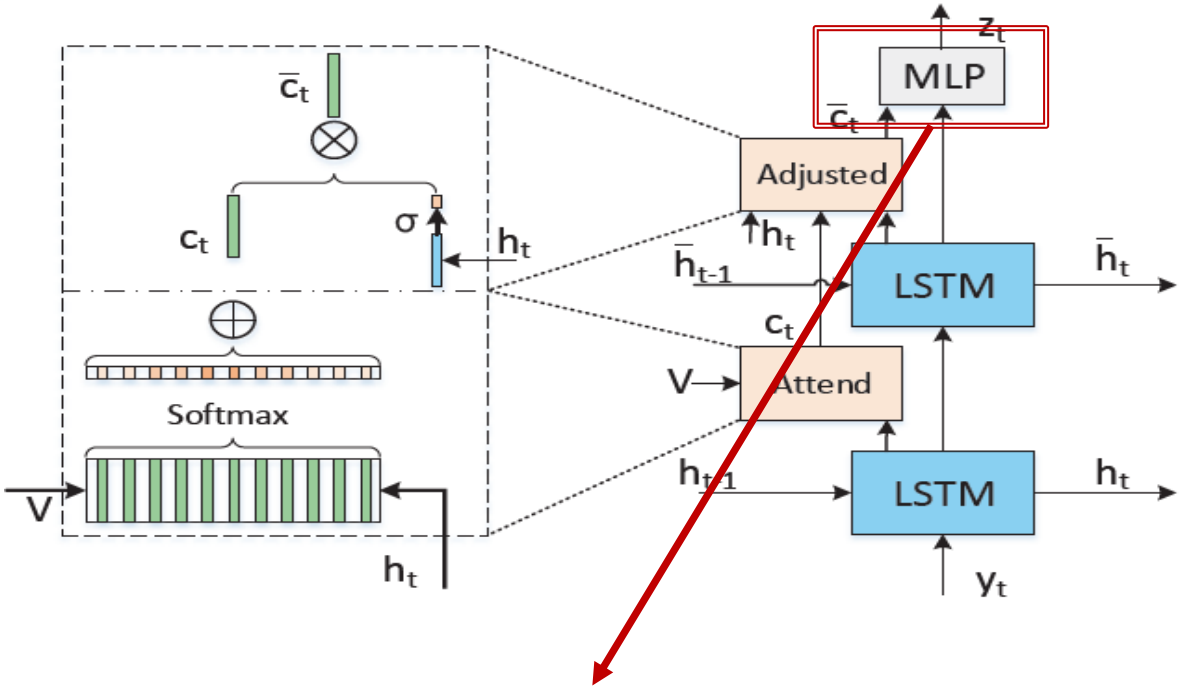


Adjusted Temporal Attention Model

decides whether **visual information or language context**

$$\bar{\mathbf{c}}_t = \beta_t \mathbf{c}_t + (1 - \beta_t)\bar{\mathbf{h}}_t$$
$$\beta_t = sigmoid(\mathbf{W}_s \mathbf{h}_t)$$

hLSTM with adjusted temporal attention



MLP Layer

outputs a predicted word with probabilities

$$\mathbf{p}_t = softmax\left(\mathbf{U}_p\phi(\mathbf{W}_p[\mathbf{h}_t;\bar{\mathbf{c}}_t]+\mathbf{b}_p)+\mathbf{d}\right)$$

Comparison study

The performance comparison with the state-of-the-art methods on MSVD. (V) denotes VGGNet, (O) denotes optical flow, (G) denotes GoogleNet, (C) denotes C3D and (R) denotes ResNet-152

Model	B@1	B@2	B@3	B@4	METEOR	CIDEr
S2VT(V) [Venugopalan *et al.*, 2015]	-	-	-	-	29.2	-
S2VT(V+O)	-	-	-	-	29.8	-
HRNE(G) [Pan *et al.*, 2015a]	78.4	66.1	55.1	43.6	32.1	-
HRNE-SA (G)	79.2	66.3	55.1	43.8	33.1	-
LSTM-E(V)[Pan *et al.*, 2015b]	74.9	60.9	50.6	40.2	29.5	-
LSTM-E(C)	75.7	62.3	52.0	41.7	29.9	-
LSTM-E(V+C)	78.8	66.0	55.4	45.3	31.0	-
p-RNN(V) [Yu *et al.*, 2016]	77.3	64.5	54.6	44.3	31.1	62.1
p-RNN(C)	79.7	67.9	57.9	47.4	30.3	53.6
p-RNN(V+C)	81.5	70.4	60.4	49.9	32.6	65.8
hLSTMt (R)	82.5	71.9	62.0	52.1	33.3	73.5
hLSTMat (R)	**82.9**	**72.2**	**63.0**	**53.0**	**33.6**	**73.8**

Comparison study

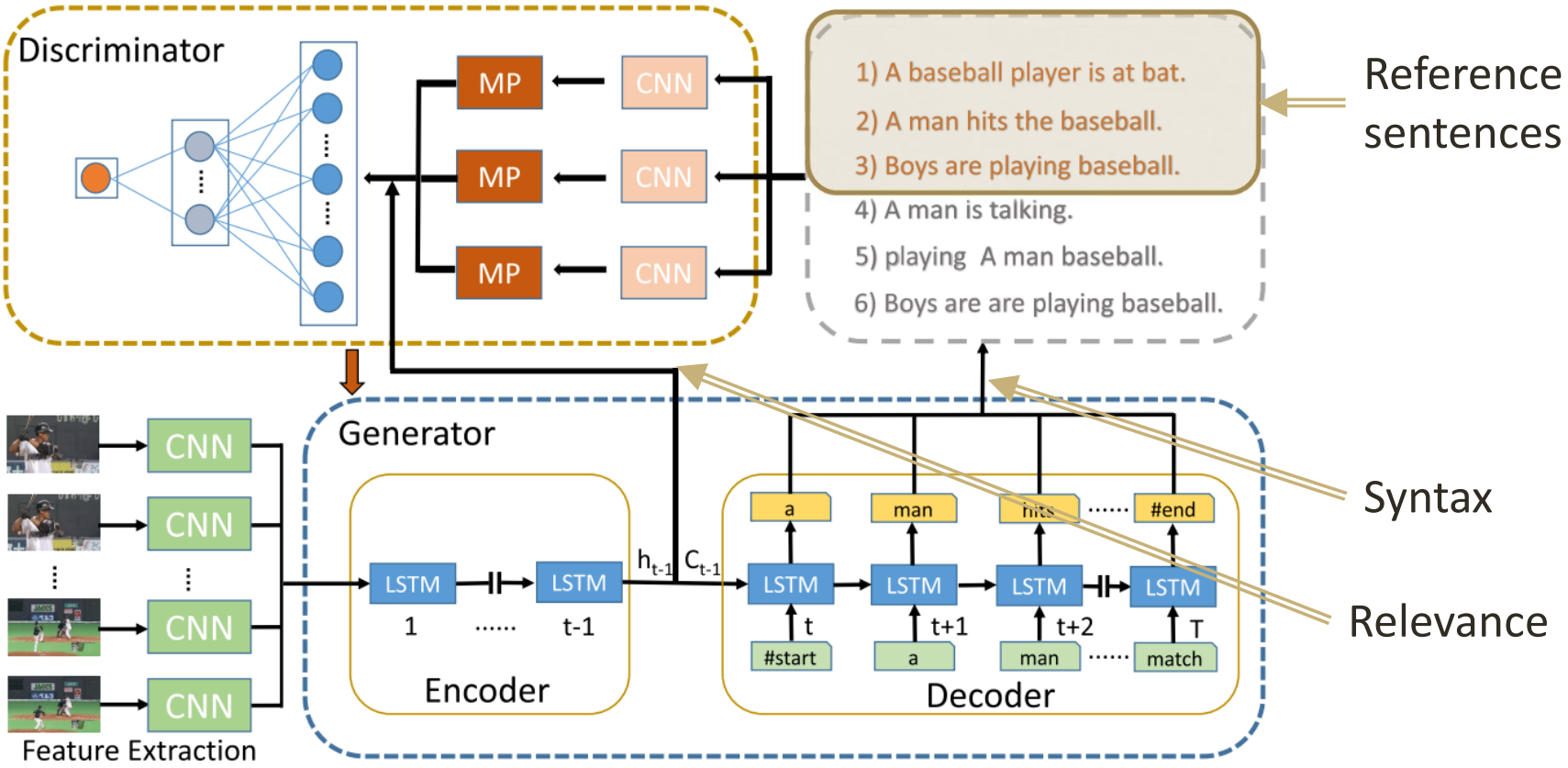
The performance comparison with the state-of-the-art methods on MSR-VTT

MP-LSTM (V)	34.8	24.8
MP-LSTM (C)	35.4	24.8
MP-LSTM (V+C)	35.8	25.3
SA (V)	35.6	25.4
SA (C)	36.1	25.7
SA (V+C)	36.6	25.9
hLSTMt (R)	**37.4**	**26.1**
hLSTMat (R)	**38.3**	**26.3**

Video captioning by adversarial LSTM (TIP 2018)

- **Motivations**
  - Existing methods often generate descriptions with syntax errors, e.g.,
    - "A boy is soccer playing"
    - "A boy is fighting with the soccer"
  - Existing methods could generate descriptions that are irrelevant to videos, e.g.,
    - "A boy is playing soccer" is confused with "a plane is flying"

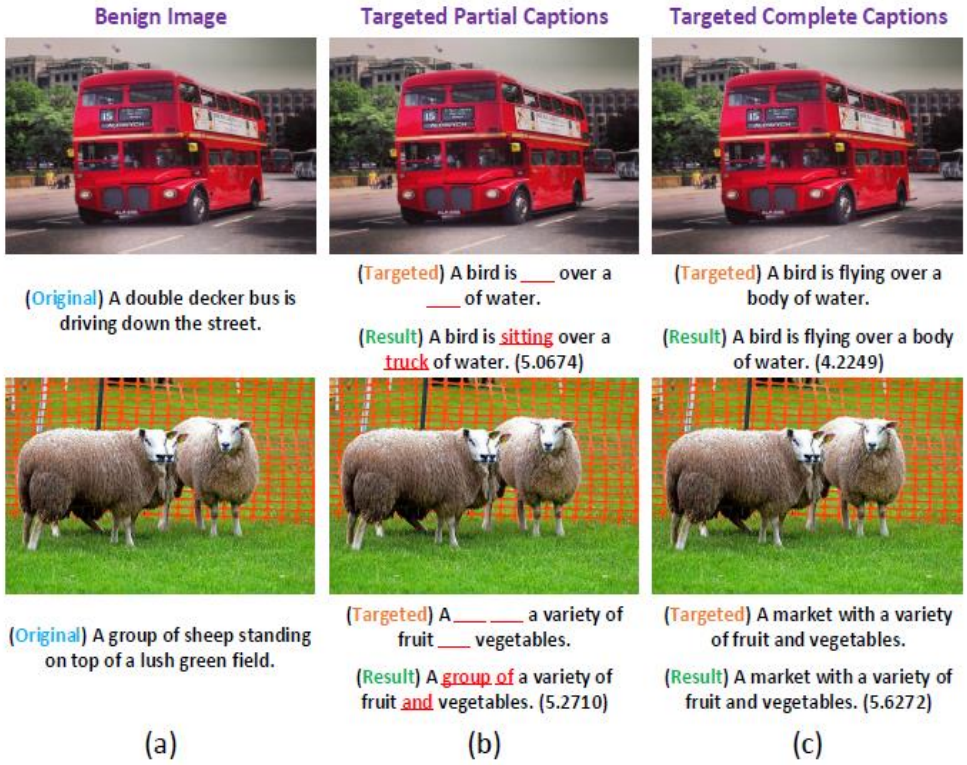
LSTM-GAN


Discriminator
MP
MP
MP
CNN
CNN
CNN
1) A baseball player is at bat.
2) A man hits the baseball.
3) Boys are playing baseball.
4) A man is talking.
5) playing A man baseball.
6) Boys are are playing baseball.
Reference sentences
Generator
CNN
CNN
CNN
CNN
Feature Extraction
LSTM
LSTM
1
t-1
Encoder
h_{t-1}
C_{t-1}
a
man
hits
#end
LSTM
LSTM
LSTM
LSTM
t
t+1
t+2
T
#start
a
man
match
Decoder
Syntax
Relevance

Adversarial attack to targeted partial captions via Structured Output Learning with Latent Variables (CVPR 2019)

Targeted partial caption:
the words at some locations are observed, while the words at other locations are not specified. (Fig. (b))

Targeted complete caption:
the words at all locations are observed. (Fig. (c))



This task has never been studied in previous work!

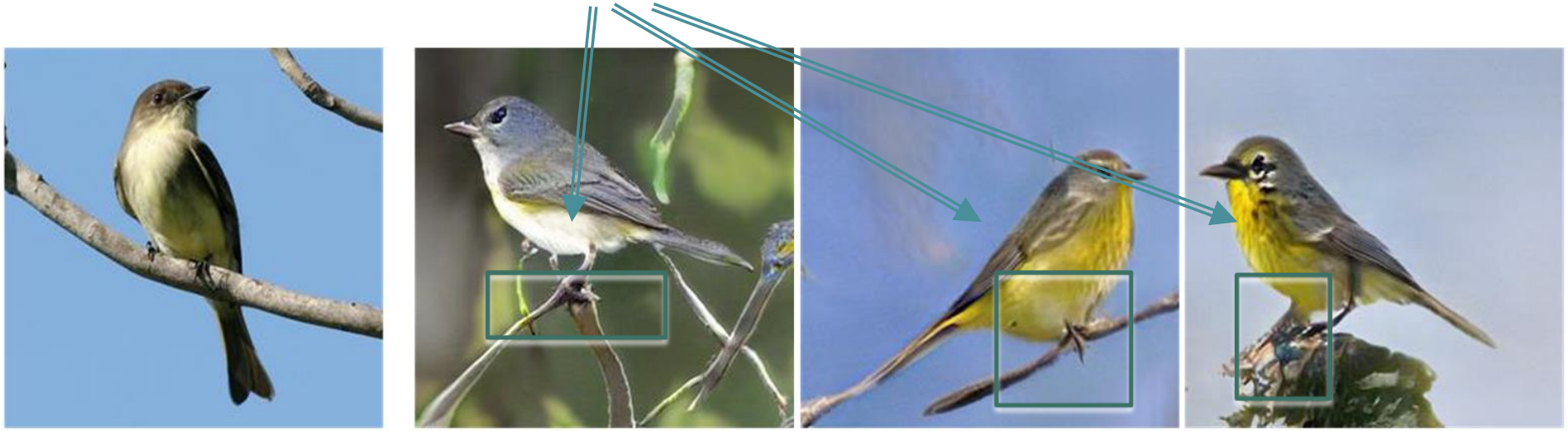
Text-to-image synthesis via Perceptual Pyramid Adversarial Networks (AAAI 2019)

Problem Statement

Generating natural images given text descriptions

Text input

A small bird with yellow and gray throat and belly, and darker crown, wings and tail feathers



Ground-truth image

Generated images

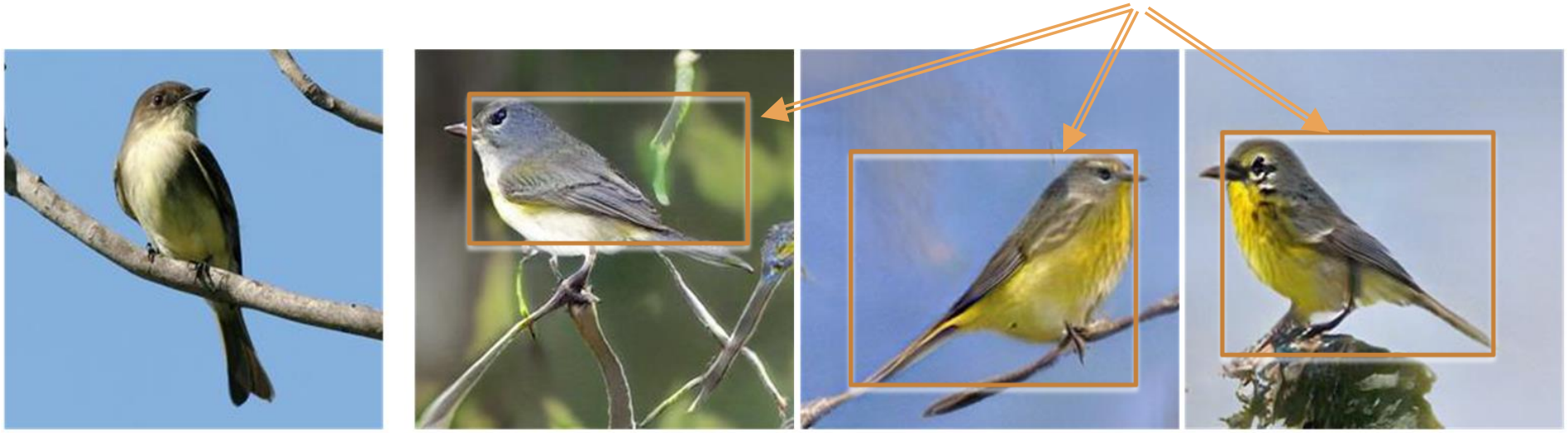
Text-to-image synthesis via Perceptual Pyramid Adversarial Networks (AAAI 2019)

Problem Statement

Generating natural images given text descriptions

Text input

A small bird with yellow and gray throat and belly, and darker crown, wings and tail feathers



Ground-truth image

Generated images

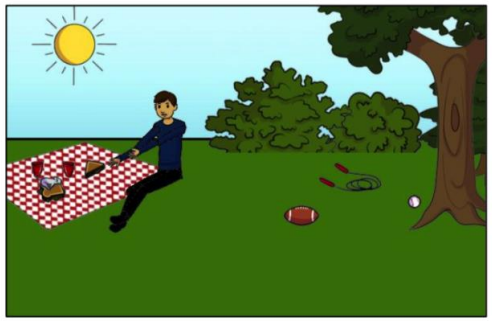
Visual Question Answering (VQA)

Problem Statement

Given an image and a natural language question about the image, providing an accurate natural language answer

It requires a potentially vast set of AI capabilities to answer

Fine Grained Recognition:
What kind of cheese is on the pizza?



Object Detection:
How many balls are there on the ground?

Commonsense Reasoning :
Does this person have 20/20 vision?

Activity Recognition:
Is the woman sitting?

Visual Question Answering (VQA)

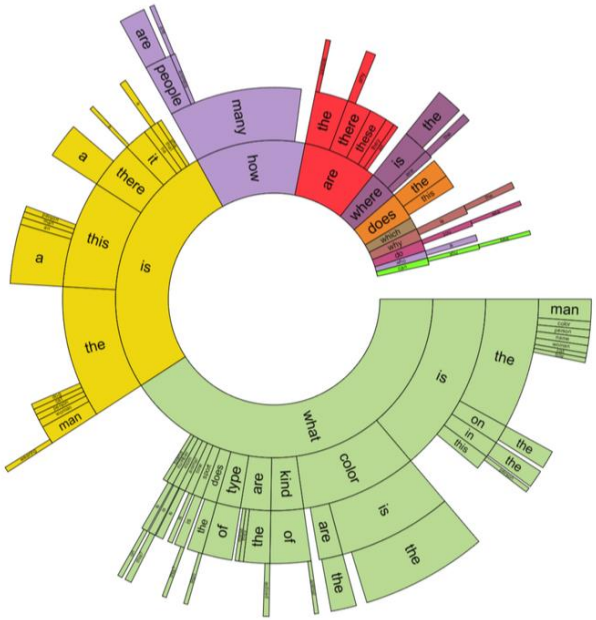
- How to enable VQA?
  - We need luxuriant mutli-modal data containing triplets of images, questions and answers
  - Selecting images with multiple objects and rich contextual information



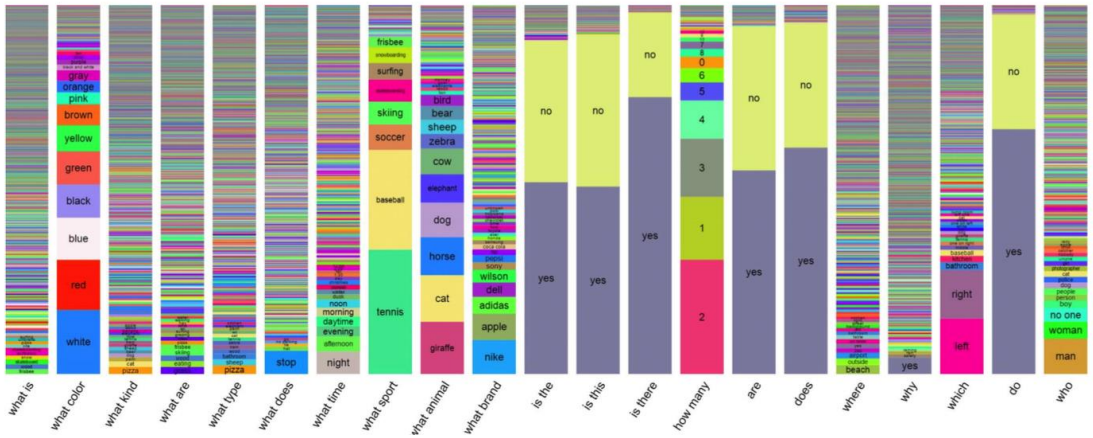

Images from MSCOCO

Visual Question Answering (VQA)

- Questions and Answers
  - User interface to collect “interesting” questions
  - Gathering various kinds of questions and answers from different workers



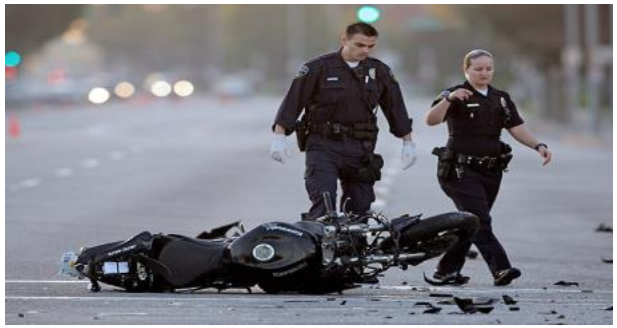
Types of Questions



Types of Answers

Visual Question Answering (VQA)

- The quality of questions is crucial for VQA
  - Quantity, Diversity, Appropriateness



Was anyone injured in the crash?

Is the motorcyclist all right?

What happened?

What caused this accident?

Was anyone injured in the crash?

Is the motorcyclist OK?

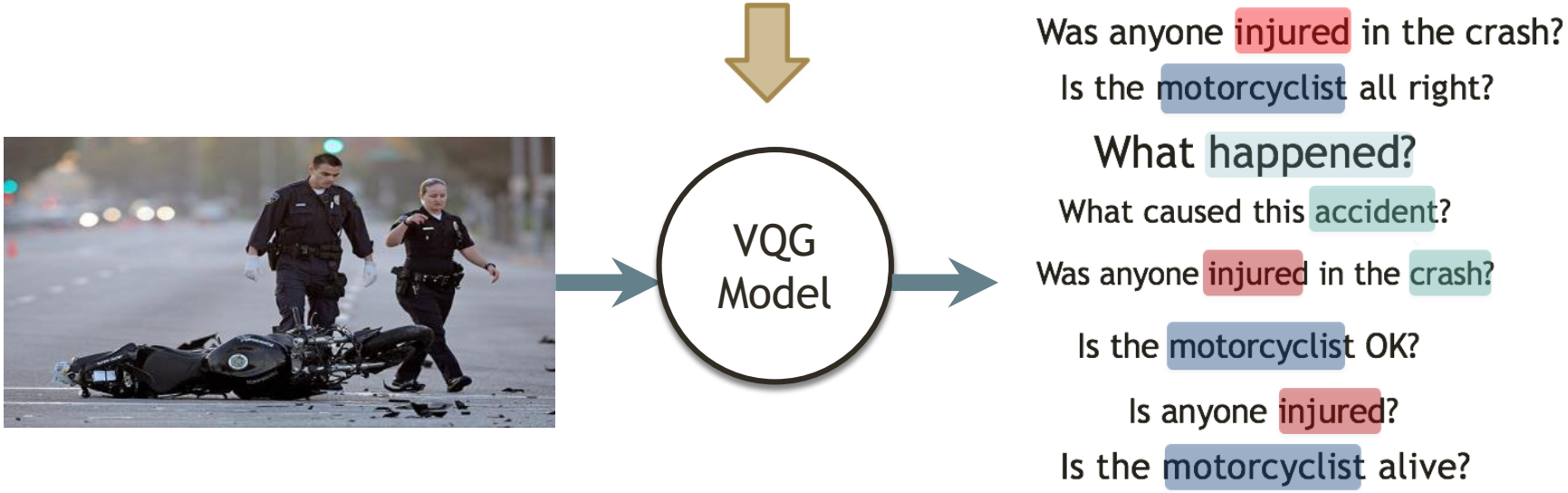
Is anyone injured?

Is the motorcyclist alive?

Manually collecting questions of high quality is costly!

Visual Question Generation (VQG)

Can we automatically generate questions for images ?



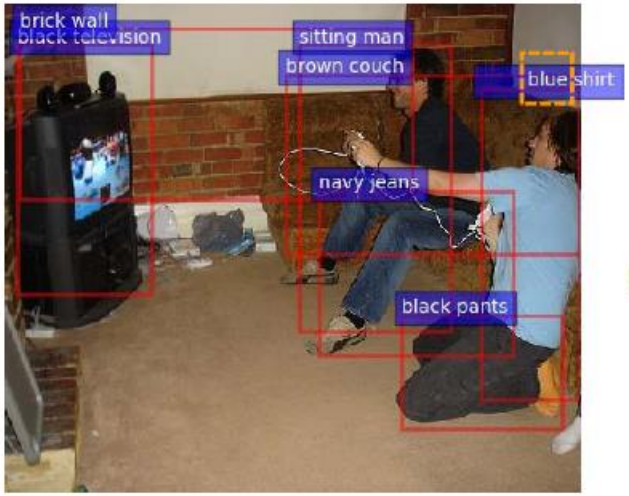
- Visual question generation (VQG)
  - It is a dual task for the VQA, which can provide infinite sources of questions for the VQA
  - For VQG, can we get some hints from VQA?

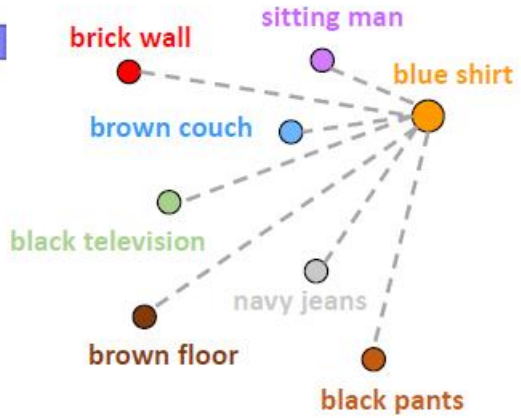
Radial Graph Convolutional Network for Visual Question Generation (TNNLS 2020)

- Rethinking the relation between "image", "question" and "answer"

The ***answer*** area that corresponds to the latent answer "blue" can be directly located and then connected with the relevant object regions around in ***image***, where their explicit interactions can be naturally encoded in a **graph structure** to generate ***question***

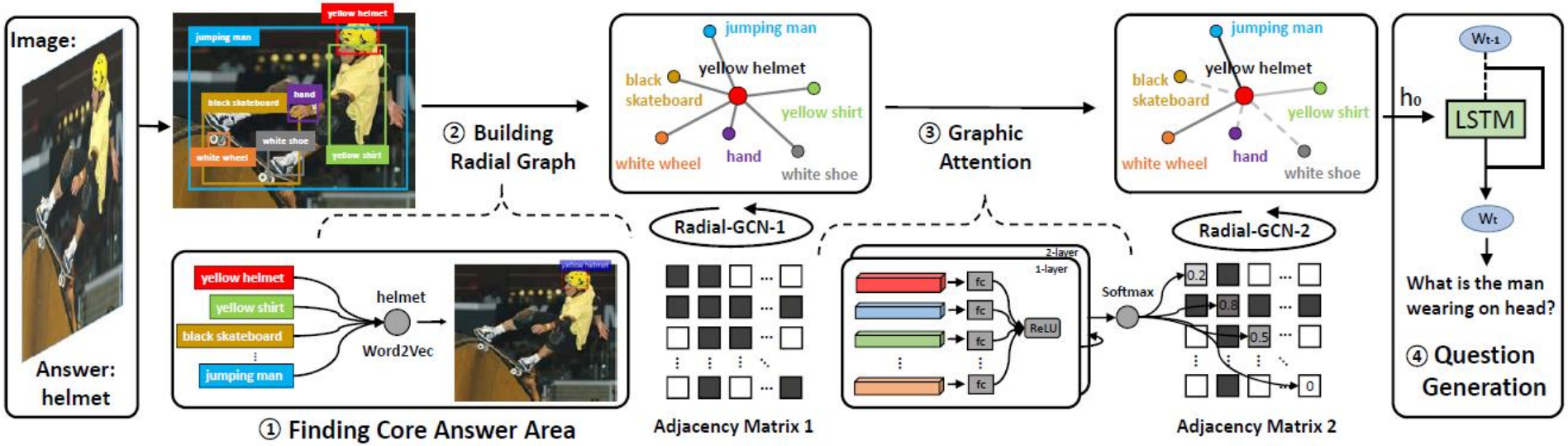
A: Blue ⟶ Q: What color is the man's shirt on the right?





Using radial graph structure to model the relation between image, answer and question

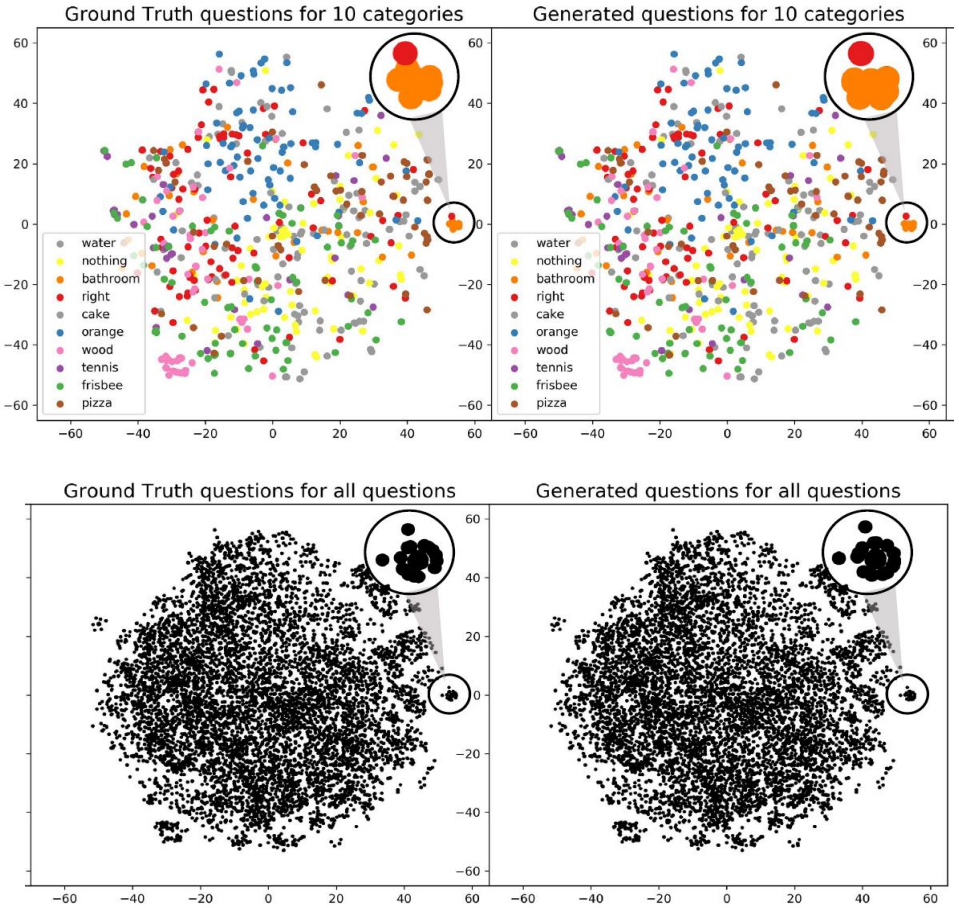
Radial Graph Convolutional Network for Visual Question Generation



- An innovative answer-centric approach focuses on the relevant image regions
- Finding the core answer area in an image by matching the latent answer with the semantic labels learned from all image regions
- Using sparse graph of radial structure is naturally built to capture the associations between the core node

Visualization

Distribution of Groundtruth questions



Distribution of Generated questions

Our proposed method can generate questions that follow the same principles and spirit of human-like questions in the ground-truth set

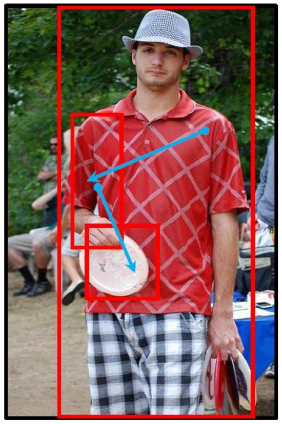
MRA-Net: Multi-modal Relation Attention Network for VQA (TPAMI 2020)

Q: What color is the T-shirt?
A: Red

(a)

Q: Is the man wearing a hat?
A: Yes

(b)

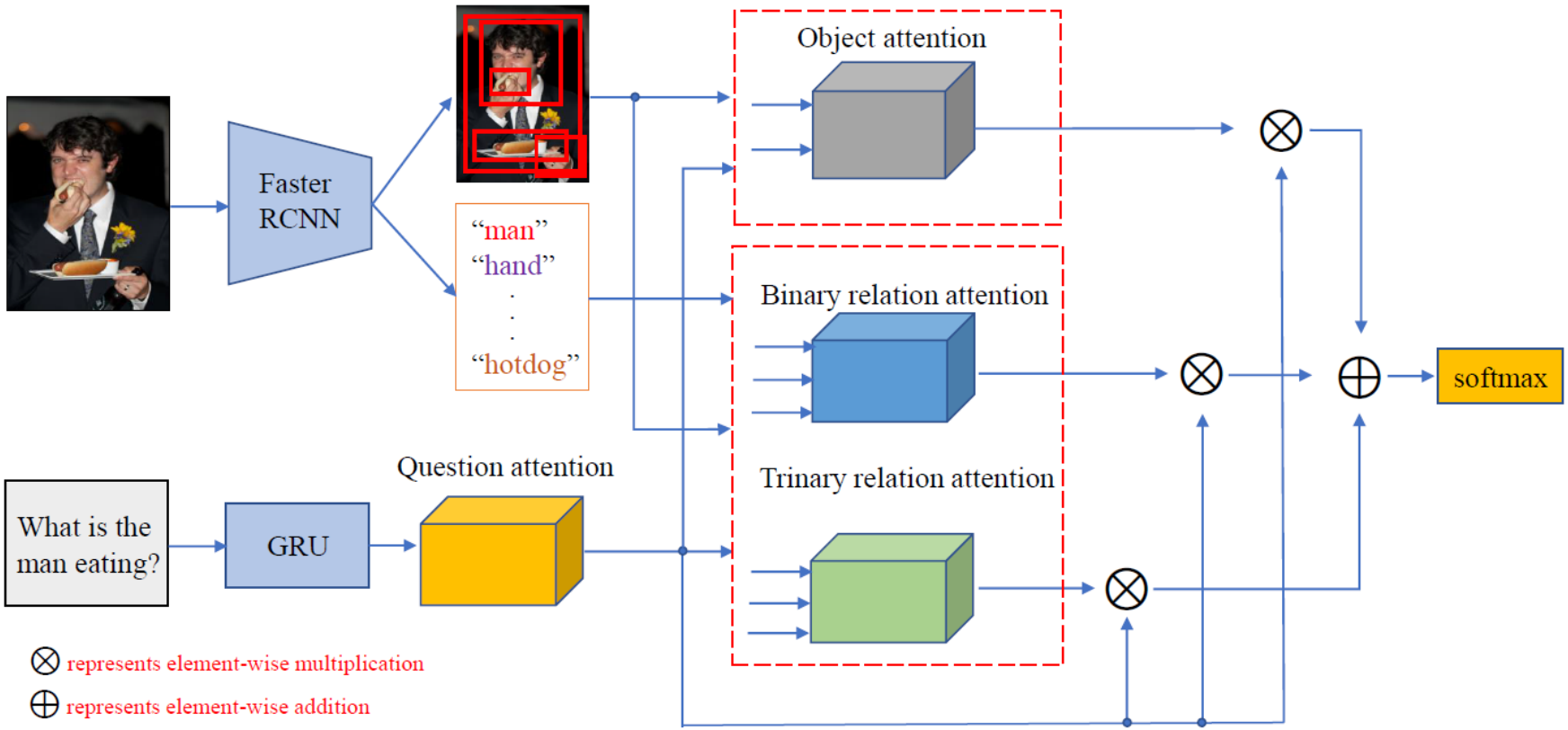


Q: What is the man's right hand holding? A: Frisbee

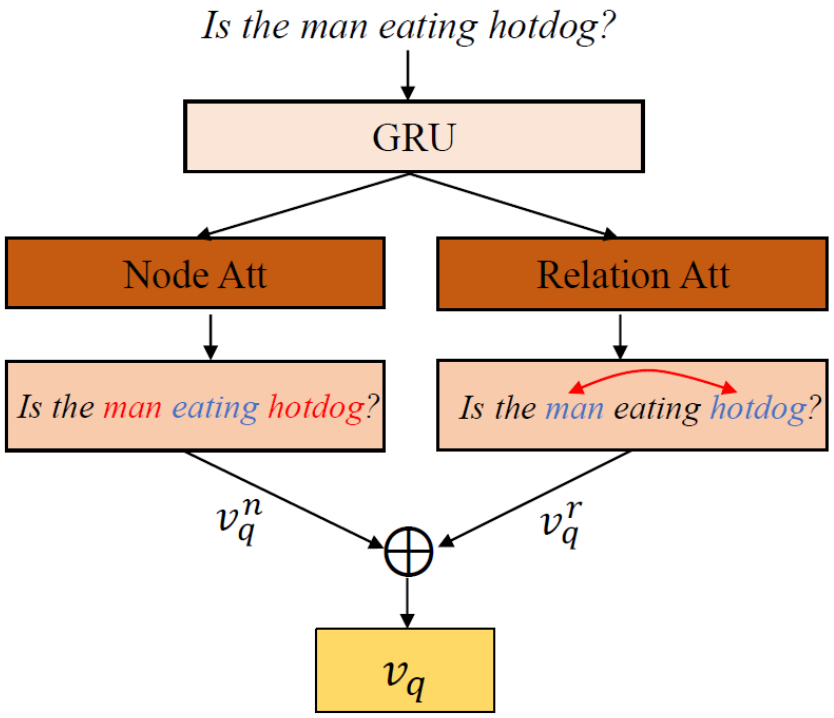
(c)

- For simple questions, object appearance features are enough
- For complicated questions, relations between objects are indispensable
- Visual relation: binary relation, trinary relation
- Relations between keywords also provide semantic relational knowledge

Overview


Faster RCNN
"man"
"hand"
"hotdog"
Object attention
Binary relation attention
Trinary relation attention
Question attention
What is the man eating?
GRU
softmax
⊗ represents element-wise multiplication
⊕ represents element-wise addition

Question Attention



- Node Attention for catching the keywords
- Relation Attention for encoding the core relationships between words

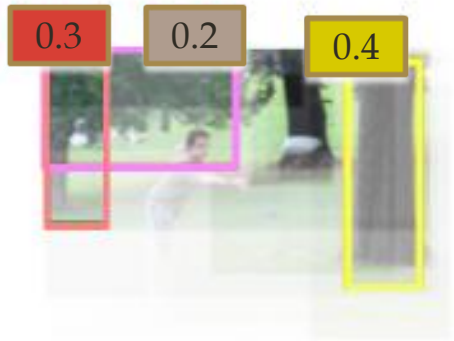
Object Attention

Q: What type of tree is the man standing close to?

$$A = \sigma(W_1^o v^q)\mathbb{1}^T \oplus \sigma(W_2^o V^o),$$

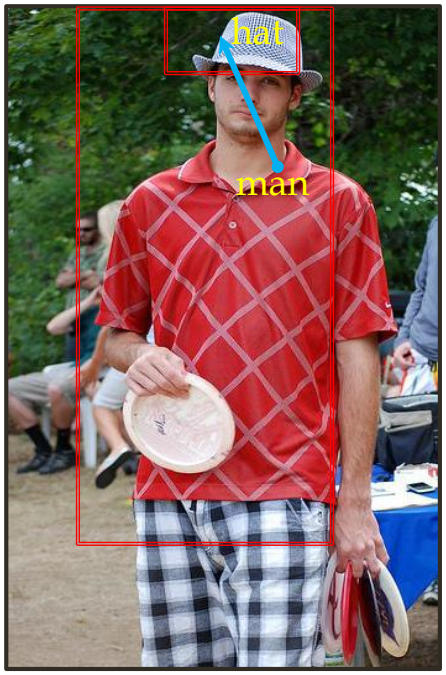
$$p^o = \text{softmax}(W_3^o A),$$

$$v^o = \sum_i p_i^o v_i,$$

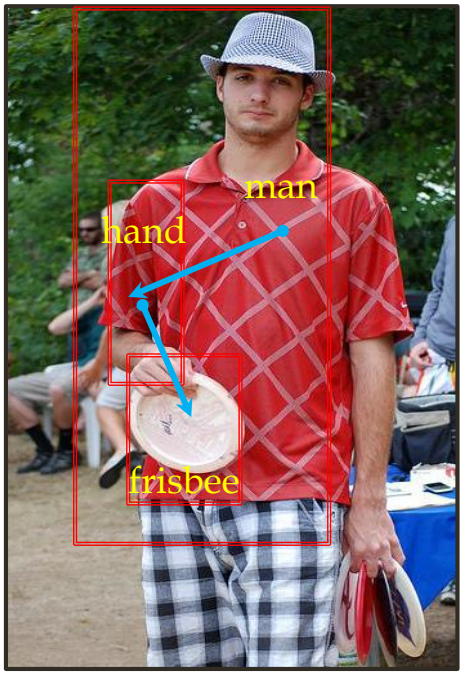


question-related objects

Visual Relation



Q: Is the man wearing a hat?
A: Yes



Q: What is the man's right hand taking? A: Frisbee

Only encode the relations of the K most important objects

$$U^{r,b}, U^{r,t} = topk(U^o \mid p^o),$$
$$C^b, C^t = topk(C \mid p^o),$$

Visual Relation

Binary Relation Attention:

Trinary Relation Attention:

Relation Embedding

$$G_e^b = \sigma(W_{1,e}^b U^{r,b} \oplus W_{2,e}^b C^b) \otimes \sigma(W_{3,e}^b v^q)\mathbb{1}^T$$

$$R_{i,j}^b = G_{1,i}^b \otimes G_{2,j}^b.$$

$$\begin{cases} p^b = \mathbf{softmax}(W_5^b R^b), \\ v^b = \sum_{i,j} p_{i,j}^b R_{i,j}^b, \end{cases}$$

Attention

$$G_e^t = \sigma(W_{1,e}^t U^{r,t} \oplus W_{2,e}^t C^t) \otimes \sigma(W_{3,e}^t v^q)\mathbb{1}^T$$

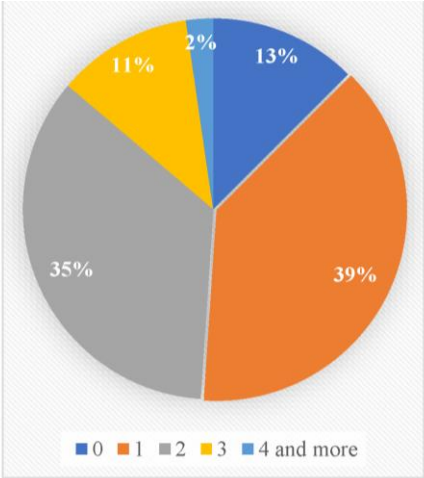
$$R_{i,j,g}^t = G_{1,i}^t \otimes G_{2,j}^t \otimes G_{3,g}^t,$$

$$\begin{cases} p^t = \mathbf{softmax}(W_7^t R^t), \\ v^t = \sum_{i,j,g} p_{i,j,g}^t R_{i,j,g}^t, \end{cases}$$

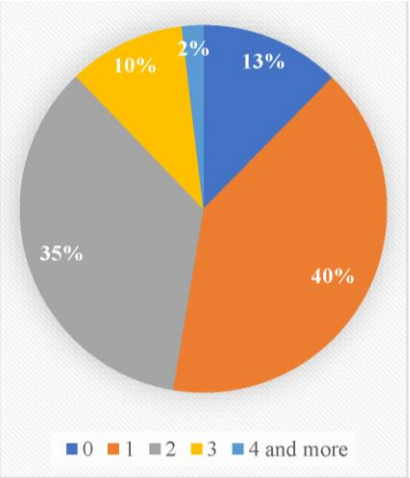
Visual Relation

Why not quaternary or more-nary?

Ratio of the questions that involve different number of objects

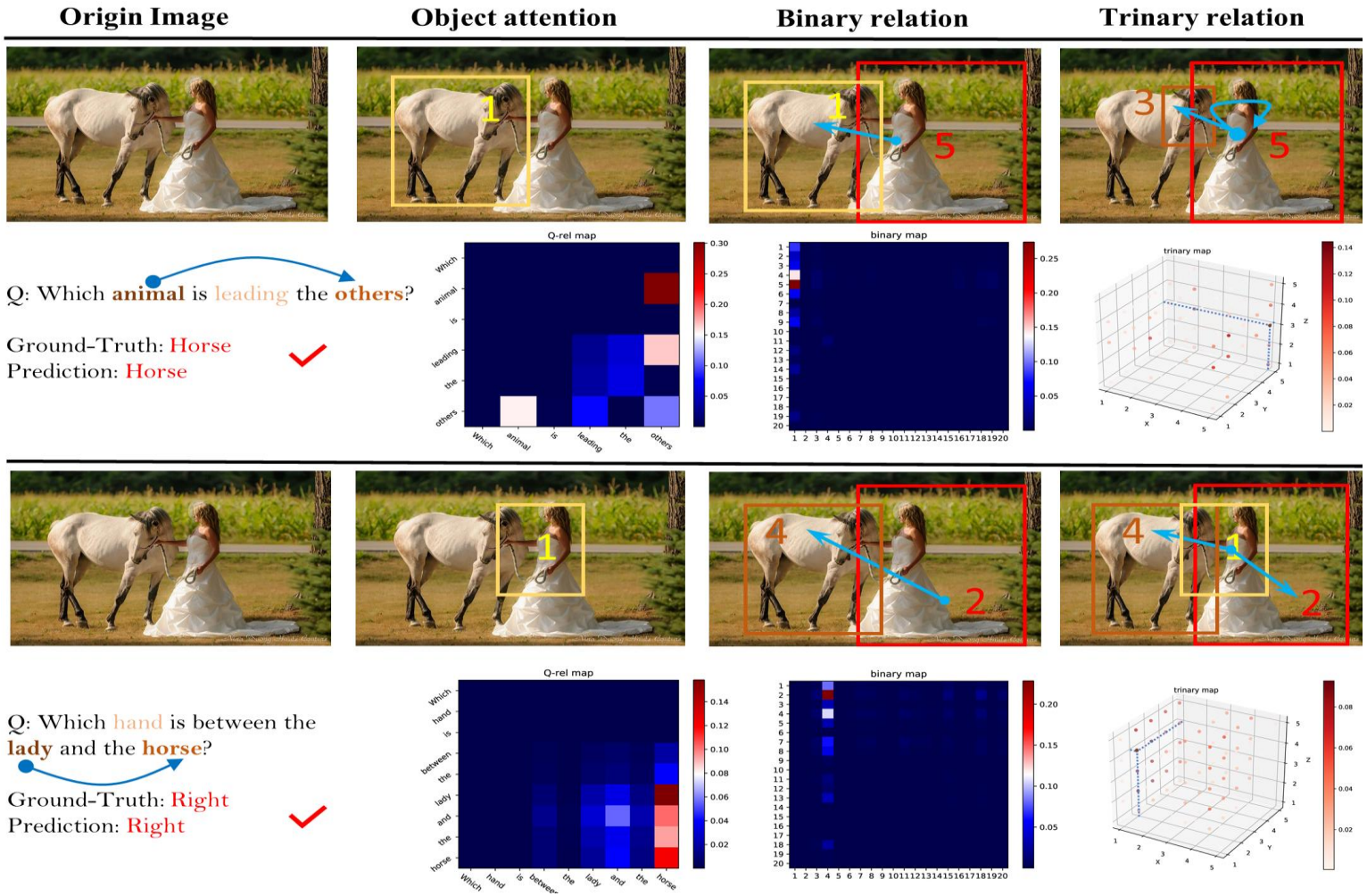


(a) Training set



(b) Testing set

Visualization



Vison and Language: the Future

- 人类大量先验知识需要被探索与结合
- 符合人类认知习惯的视觉认知场景有待探索

亟待研究协同视觉与语言处理的视觉自然认知技术

Vison and Language: the Future

视觉自然认知

协同视觉与语言处理的

跨模态数据特征提取

普适的深度神经网络，能够并行与快速的同时提取视觉与语言特征

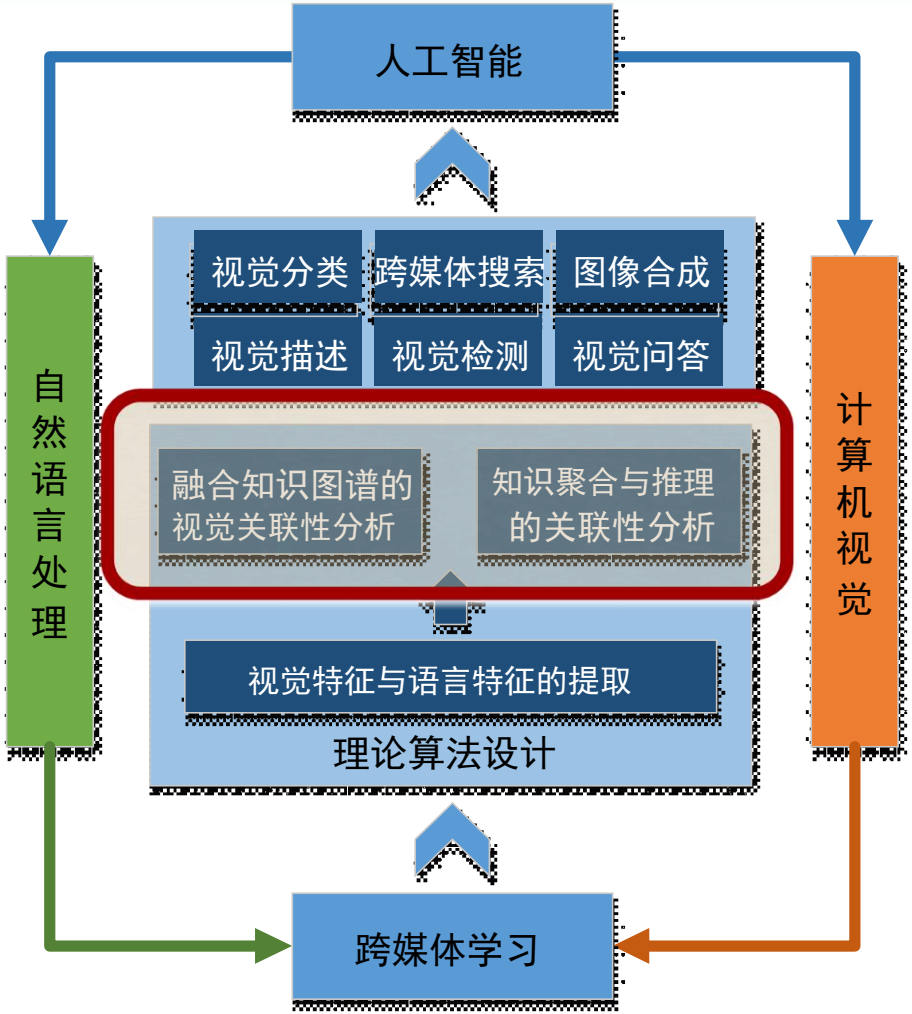
关联性分析

将视觉信息进行结构化的表达，让模型能够基于结构化知识图谱对自然知识进行聚合和推理

部署应用

基于视觉和语言的关联性分析，将其部署在有实用价值的应用上，提供跨模态解决方案

Vison and Language: the Future



Thank you!

shenhengtao@hotmail.com

未来媒体研究中心
CENTER FOR FUTURE MEDIA

电子科技大学
University of Electronic Science and Technology of China