COMP 786 (Fall 2020)
Natural Language Processing

Week 11: Language+Vision (incl. several Guest Research Talks)

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Mohit Bansal

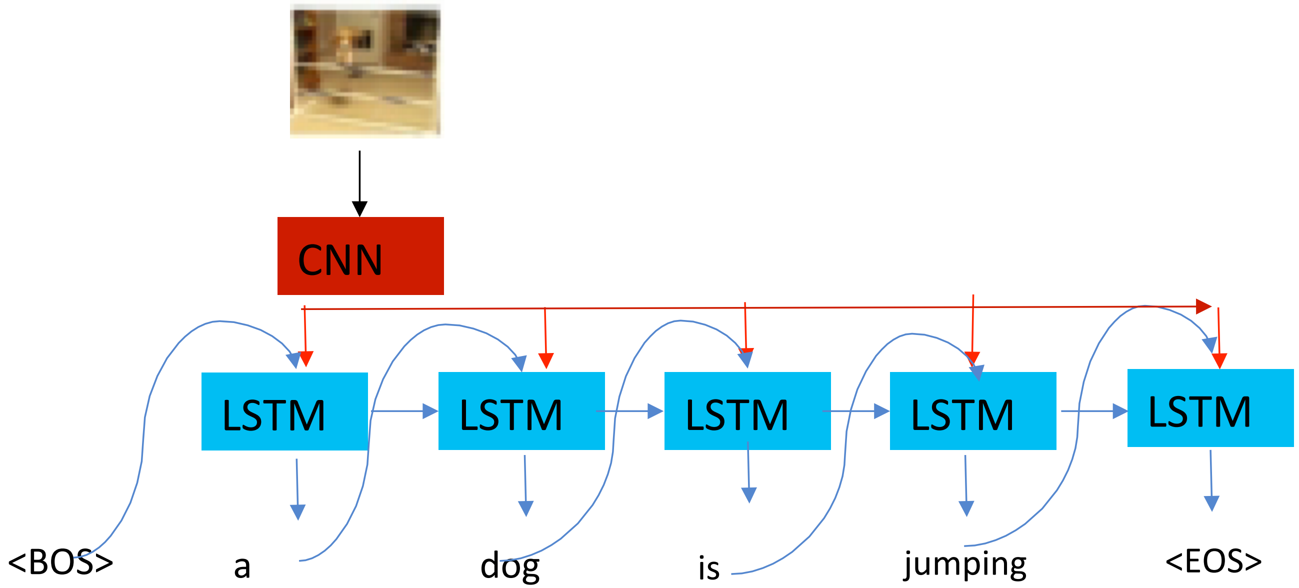
Language+Vision

Example Major Language+Vision Tasks

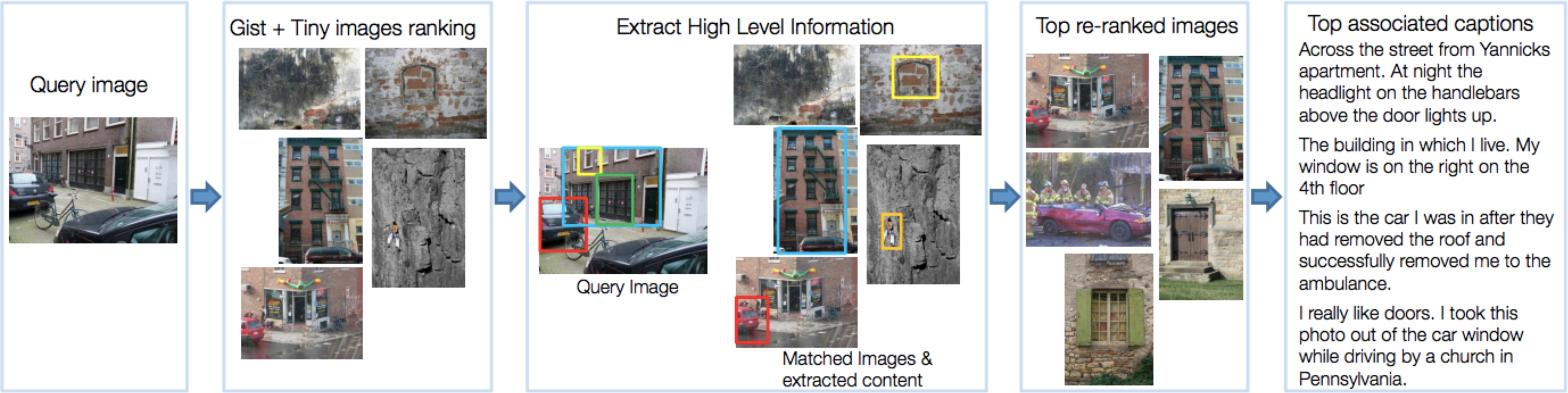
- Image Captioning
- Referring Expressions
- Image/Visual Question Answering
- Visual Dialog
- Video Captioning
- Video QA/Dialogue
- Cross-Modal Pretraining Models & Text-to-Image Generation

Brief Task Definitions and Example Papers/Models

Image Captioning

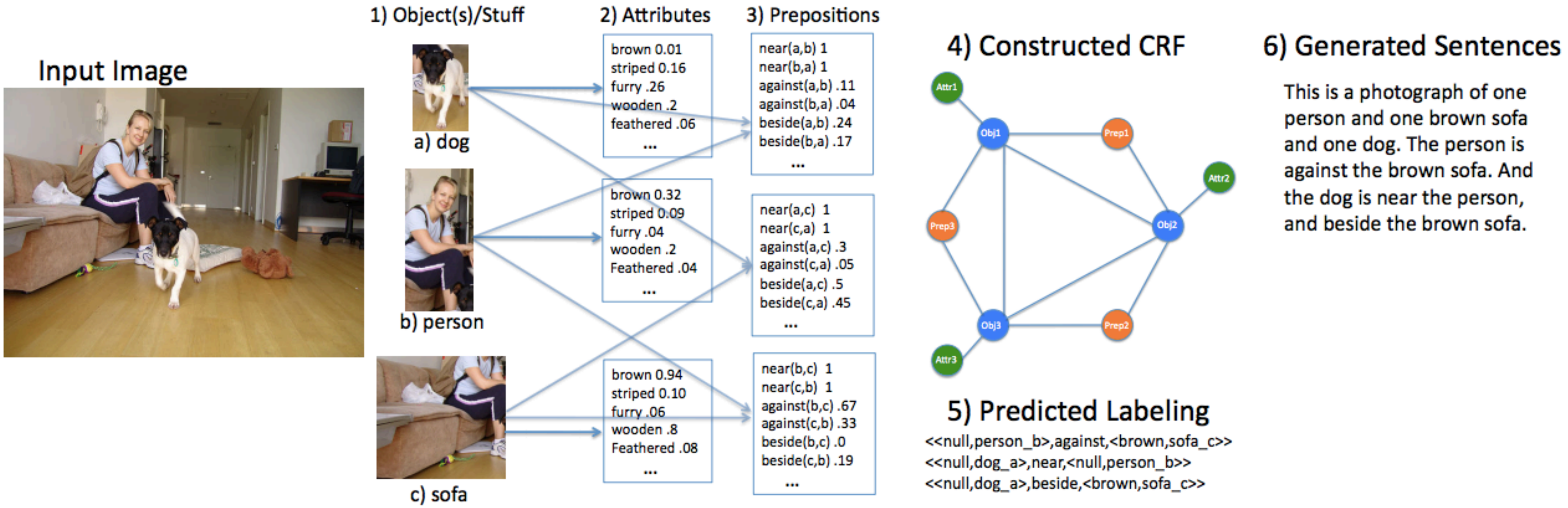


Example Early Methods



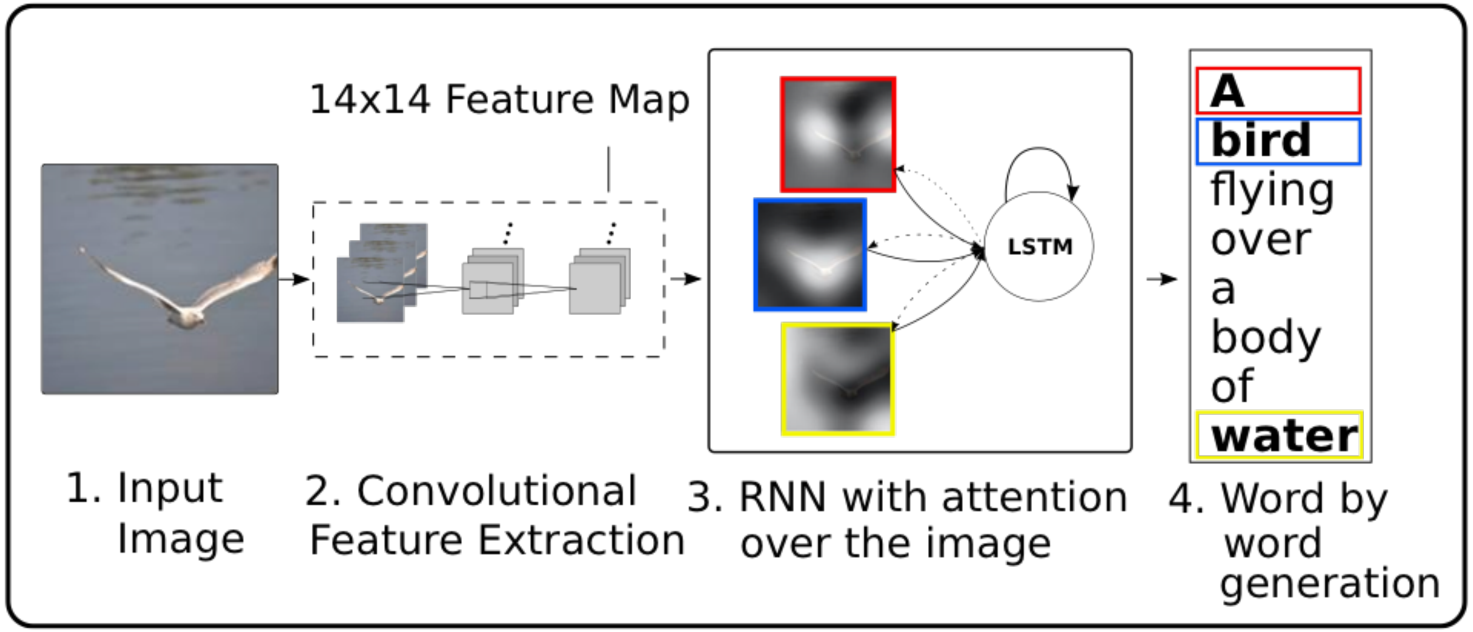
[Ordonez et al., 2011]

Example Early Methods



[Kulkarni et al., 2013]

Show, Attend, and Tell



Attention:

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.$$

[Xu et al., 2015]

Show, Attend, and Tell

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

[Xu et al., 2015]

Visual Referring Expressions

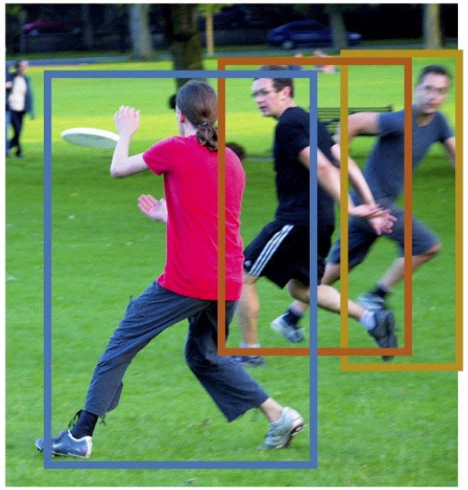
RefCOCO TestB

bottom left banana

second banana from left

top right banana

RefCOCO+ TestA

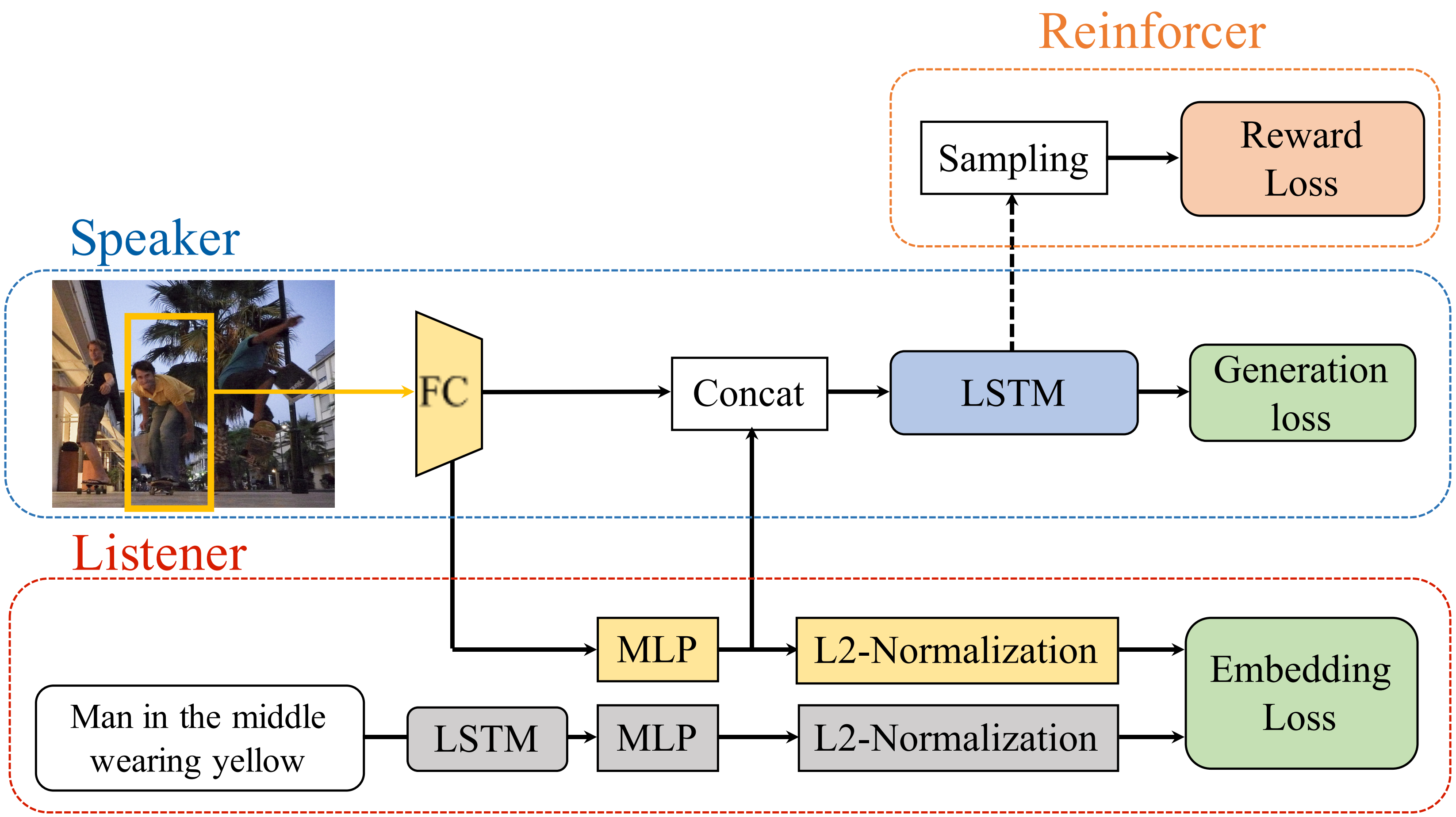


red shirt

man in black

blue shirt

Joint Comprehension+Generation Model



[Yu et al., 2017]

Joint Comprehension+Generation Model

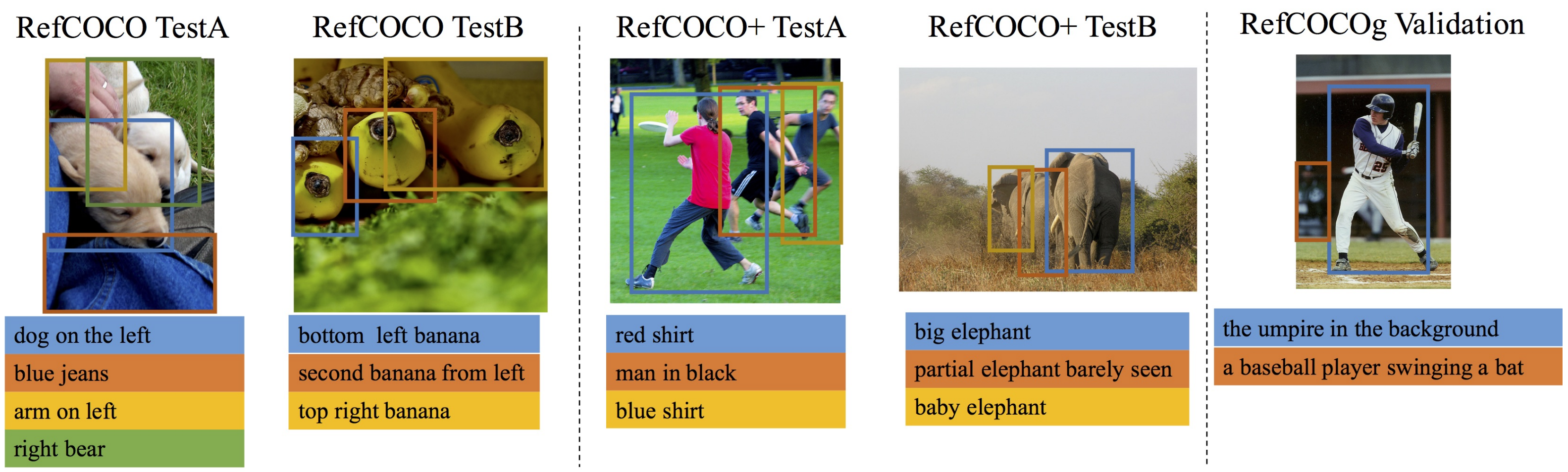


Figure 1: Joint generation examples using our full model with "+rerank" on three datasets. Each sentence shows the generated expression for one of the depicted objects (color coded to indicate correspondence).

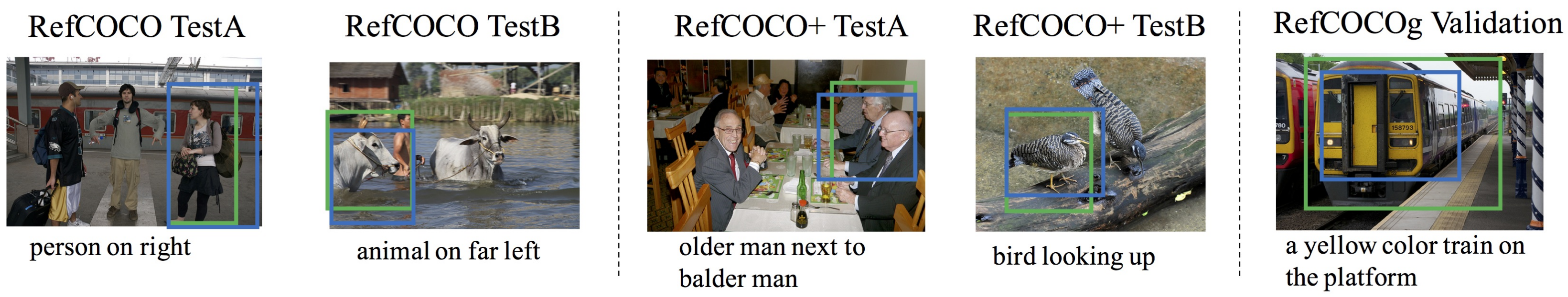


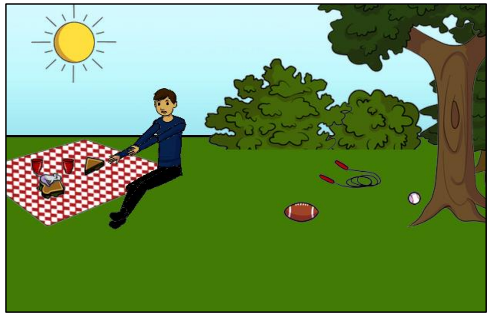
Figure 2: Example comprehension results using our full model on three datasets. Green box shows the ground-truth region and blue box shows our correct comprehension based on the detected regions.

[Yu et al., 2017]

VQA: Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



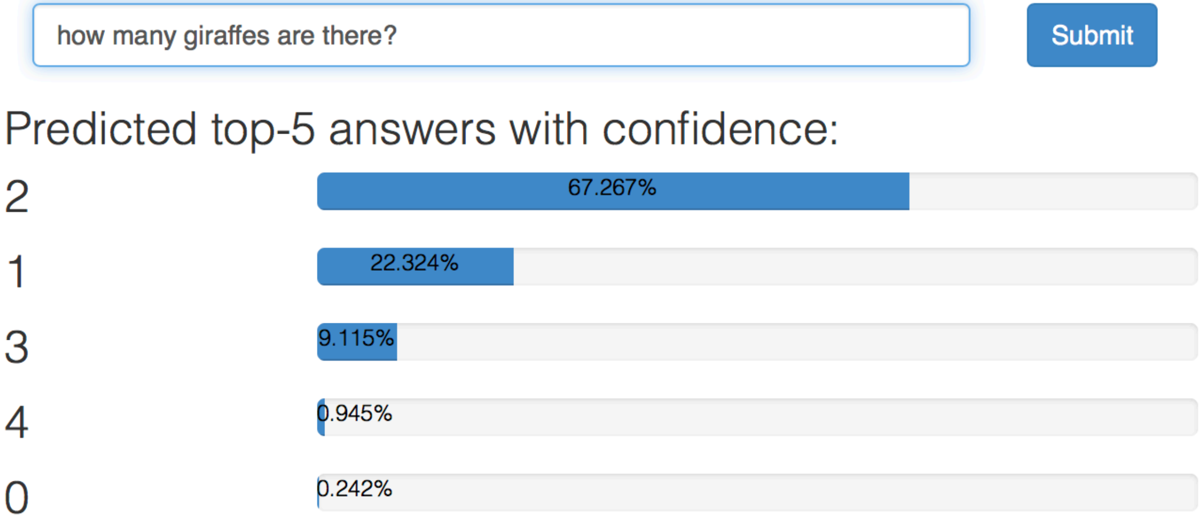
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

[Agrawal et al., 2015]

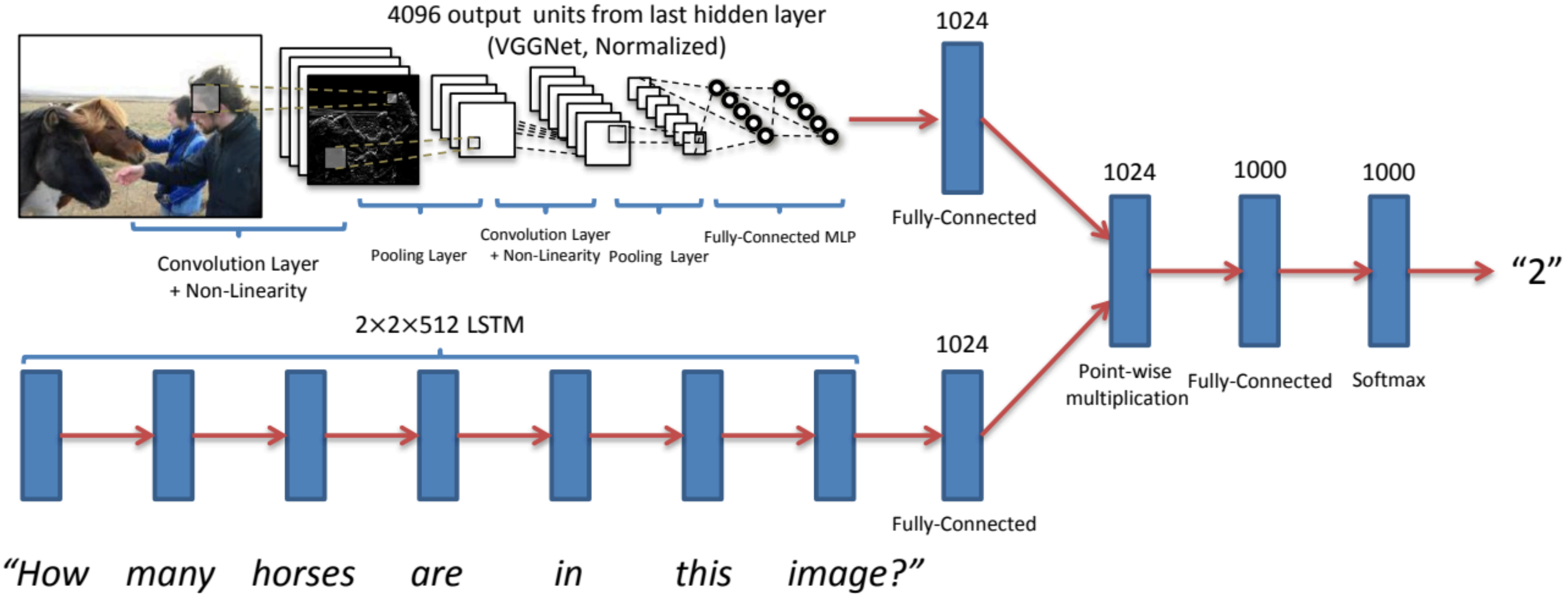
Demo

- http://vqa.cloudcv.org/



[Agrawal et al., 2015]

Simple VQA Baseline



[Agrawal et al., 2015]

Hierarchical Co-Attention Model

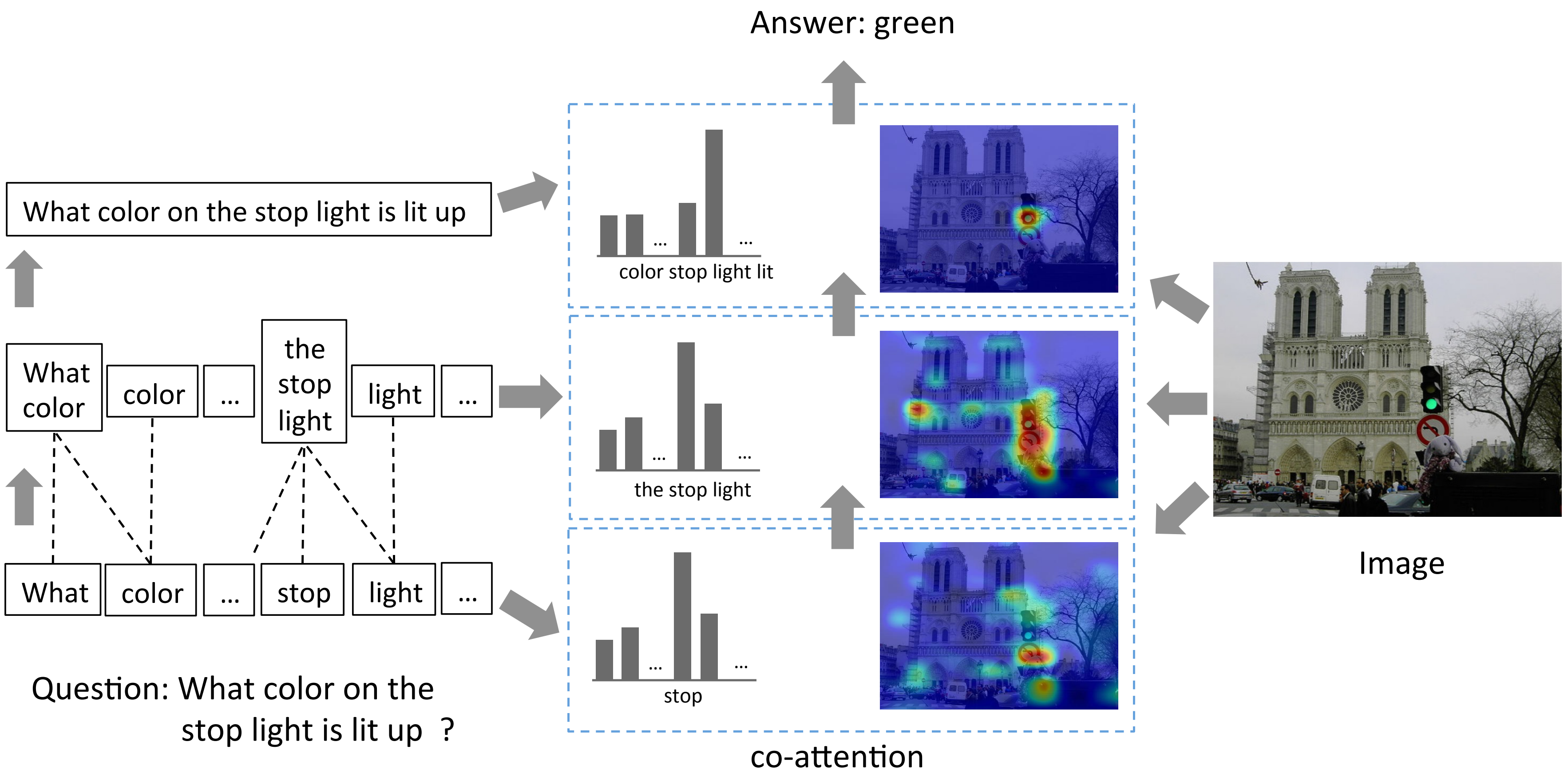


Figure 1: Flowchart of our proposed hierarchical co-attention model. Given a question, we extract its word level, phrase level and question level embeddings. At each level, we apply co-attention on both the image and question. The final answer prediction is based on all the co-attended image and question features.

[Lu et al., 2016]

Hierarchical Co-Attention Model

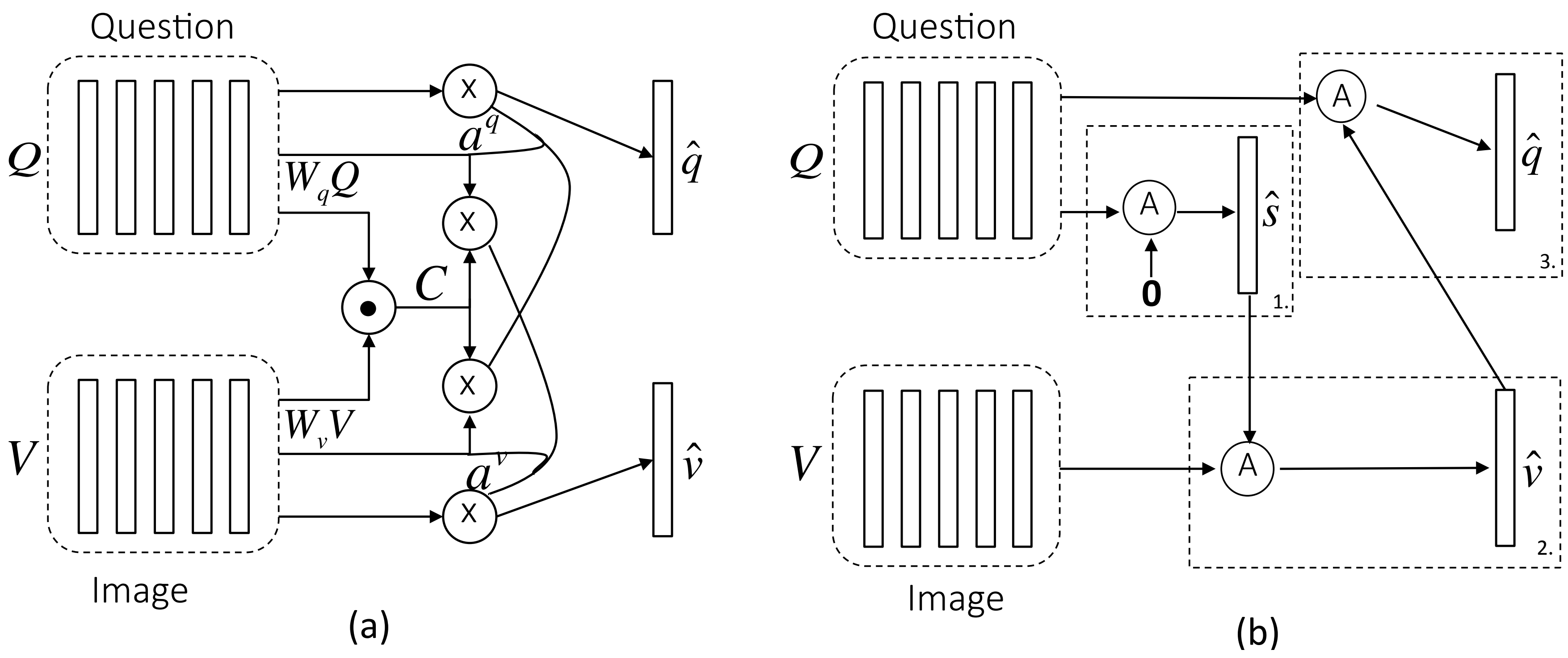


Figure 2: (a) Parallel co-attention mechanism; (b) Alternating co-attention mechanism.

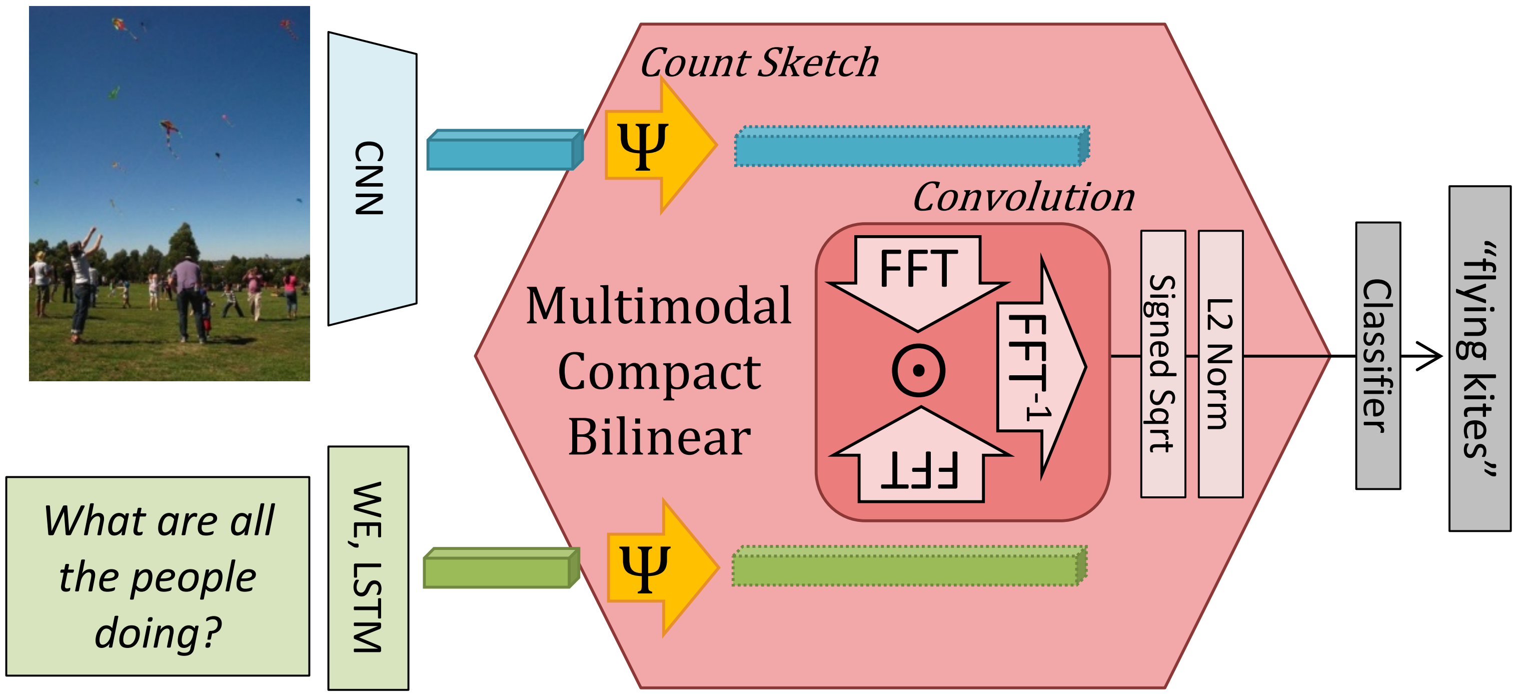
[Lu et al., 2016]

Hierarchical Co-Attention Model

Figure 4: Visualization of image and question co-attention maps on the COCO-QA dataset. From left to right: original image and question pairs, word level co-attention maps, phrase level co-attention maps and question level co-attention maps. For visualization, both image and question attentions are scaled (from red:high to blue:low). Best viewed in color.

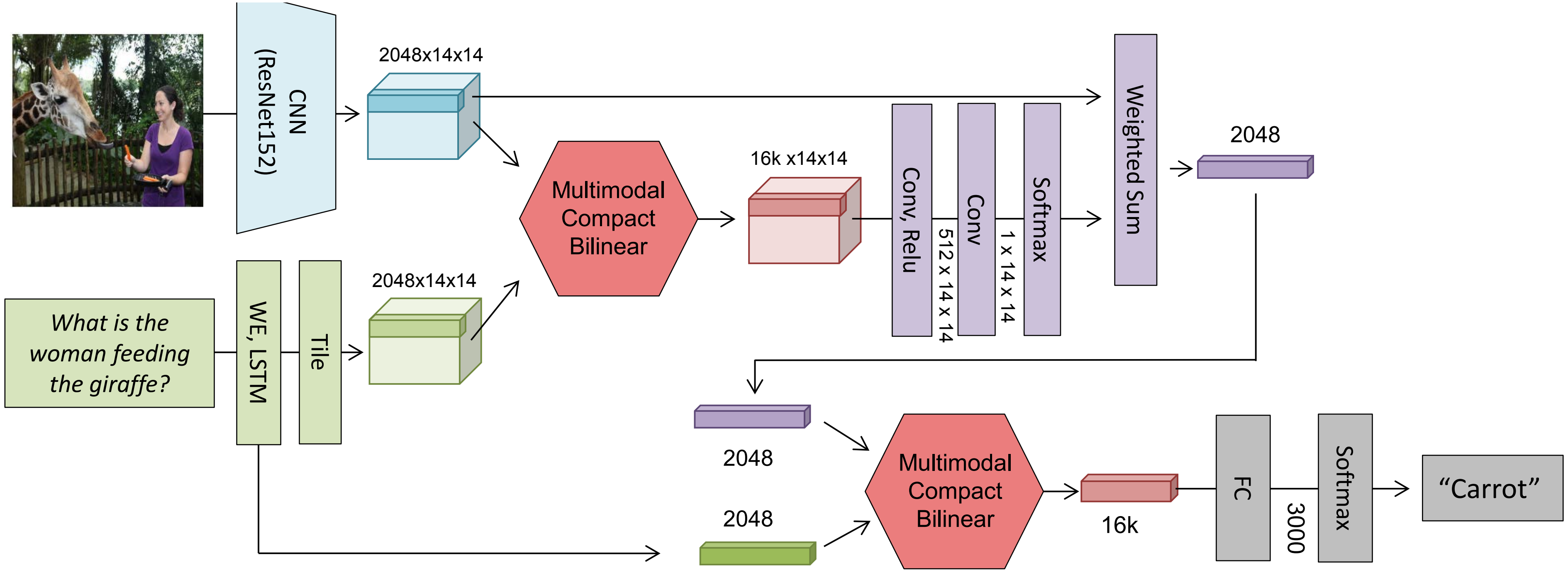
[Lu et al., 2016]

Multimodal Compact Bilinear (MCB) Model



[Fukui et al., 2016]

MCB Model with Attention



[Fukui et al., 2016]

Results

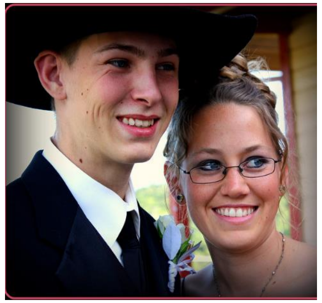
	Test-dev					Test-standard				
	Open Ended				MC	Open Ended				MC
	Y/N	No.	Other	All	All	Y/N	No.	Other	All	All
MCB	81.2	35.1	49.3	60.8	65.4	-	-	-	-	-
MCB + Genome	81.7	36.6	51.5	62.3	66.4	-	-	-	-	-
MCB + Att.	82.2	37.7	54.8	64.2	68.6	-	-	-	-	-
MCB + Att. + GloVe	82.5	37.6	55.6	64.7	69.1	-	-	-	-	-
MCB + Att. + Genome	81.7	38.2	57.0	65.1	69.5	-	-	-	-	-
MCB + Att. + GloVe + Genome	82.3	37.2	57.4	65.4	69.9	-	-	-	-	-
Ensemble of 7 Att. models	**83.4**	**39.8**	**58.5**	**66.7**	**70.2**	**83.2**	**39.5**	**58.0**	**66.5**	**70.1**
Naver Labs (challenge 2nd)	83.5	39.8	54.8	64.9	69.4	83.3	38.7	54.6	64.8	69.3
HieCoAtt (Lu et al., 2016)	79.7	38.7	51.7	61.8	65.8	-	-	-	62.1	66.1
DMN+ (Xiong et al., 2016)	80.5	36.8	48.3	60.3	-	-	-	-	60.4	-
FDA (Ilievski et al., 2016)	81.1	36.2	45.8	59.2	-	-	-	-	59.5	-
D-NMN (Andreas et al., 2016a)	81.1	38.6	45.5	59.4	-	-	-	-	59.4	-
AMA (Wu et al., 2016)	81.0	38.4	45.2	59.2	-	81.1	37.1	45.8	59.4	-
SAN (Yang et al., 2015)	79.3	36.6	46.1	58.7	-	-	-	-	58.9	-
NMN (Andreas et al., 2016b)	81.2	38.0	44.0	58.6	-	81.2	37.7	44.0	58.7	-
AYN (Malinowski et al., 2016)	78.4	36.4	46.3	58.4	-	78.2	36.3	46.3	58.4	-
SMem (Xu and Saenko, 2016)	80.9	37.3	43.1	58.0	-	80.9	37.5	43.5	58.2	-
VQA team (Antol et al., 2015)	80.5	36.8	43.1	57.8	62.7	80.6	36.5	43.7	58.2	63.1
DPPnet (Noh et al., 2015)	80.7	37.2	41.7	57.2	-	80.3	36.9	42.2	57.4	-
iBOWIMG (Zhou et al., 2015)	76.5	35.0	42.6	55.7	-	76.8	35.0	42.6	55.9	62.0

[Fukui et al., 2016]

Making the V in the VQA matter !

Who is wearing glasses?

man | woman

Where is the child sitting?

fridge | arms


Is the umbrella upside down?

yes | no


How many children are in the bed?

2 | 1


[Goyal et al., 2017]

Results

Approach	Ans Type	UU	UB	$B_{half}B$	BB
MCB [9]	Yes/No	81.20	70.40	74.89	77.37
	Number	34.80	31.61	34.69	36.66
	Other	51.19	47.90	47.43	51.23
	All	60.36	54.22	56.08	59.14
HieCoAtt [25]	Yes/No	79.99	67.62	70.93	71.80
	Number	34.83	32.12	34.07	36.53
	Other	45.55	41.96	42.11	46.25
	All	57.09	50.31	51.88	54.57

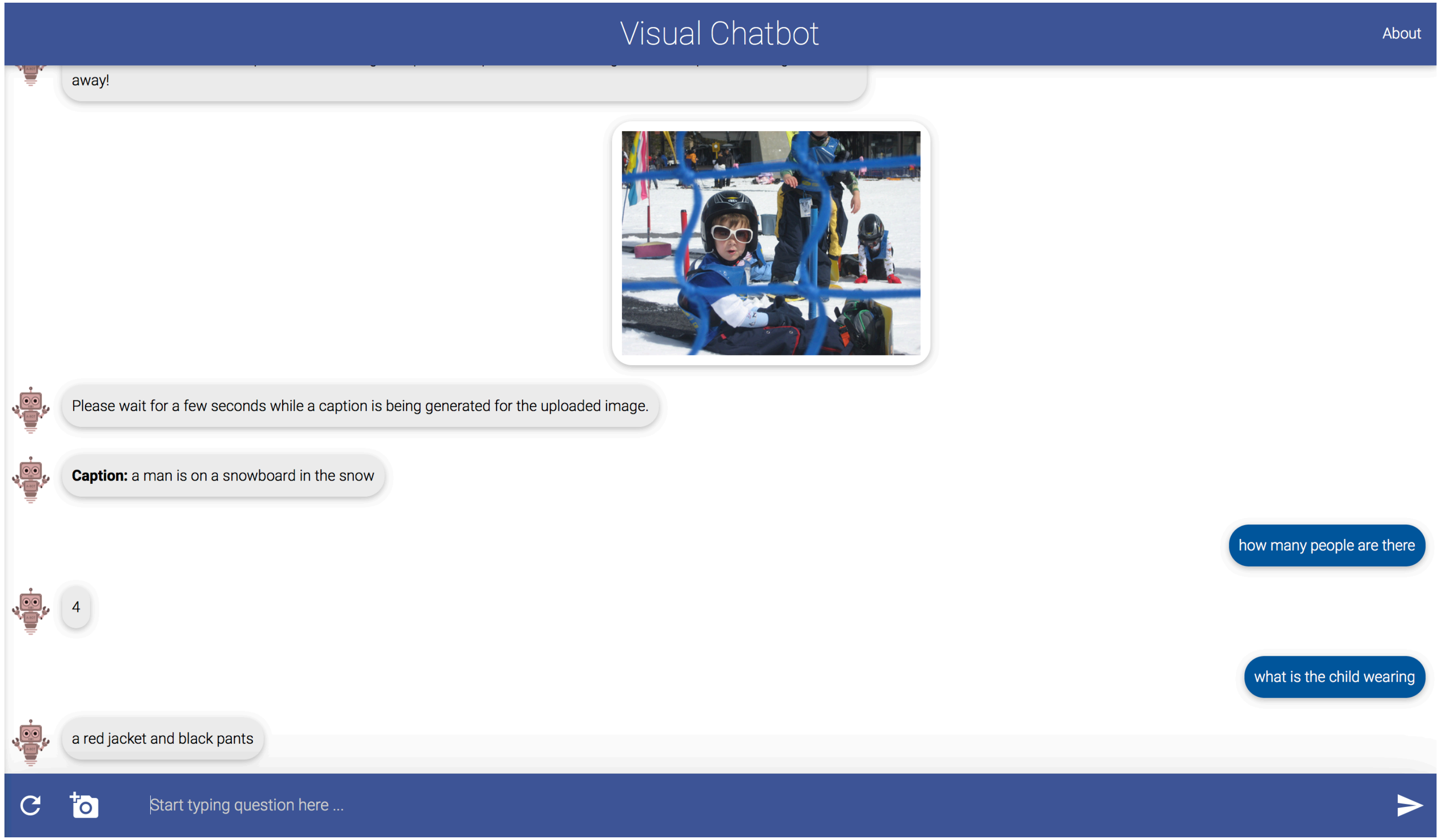
[Goyal et al., 2017]

Visual Dialog

[Das et al., 2017]

Demo

- http://visualchatbot.cloudcv.org/



[Das et al., 2017]

Visual Dialog vs VQA

VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Captioning

Two people are in a wheelchair and one is holding a racket.

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman

Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

A: Playing a Wii game

Q: Is that a man to her right

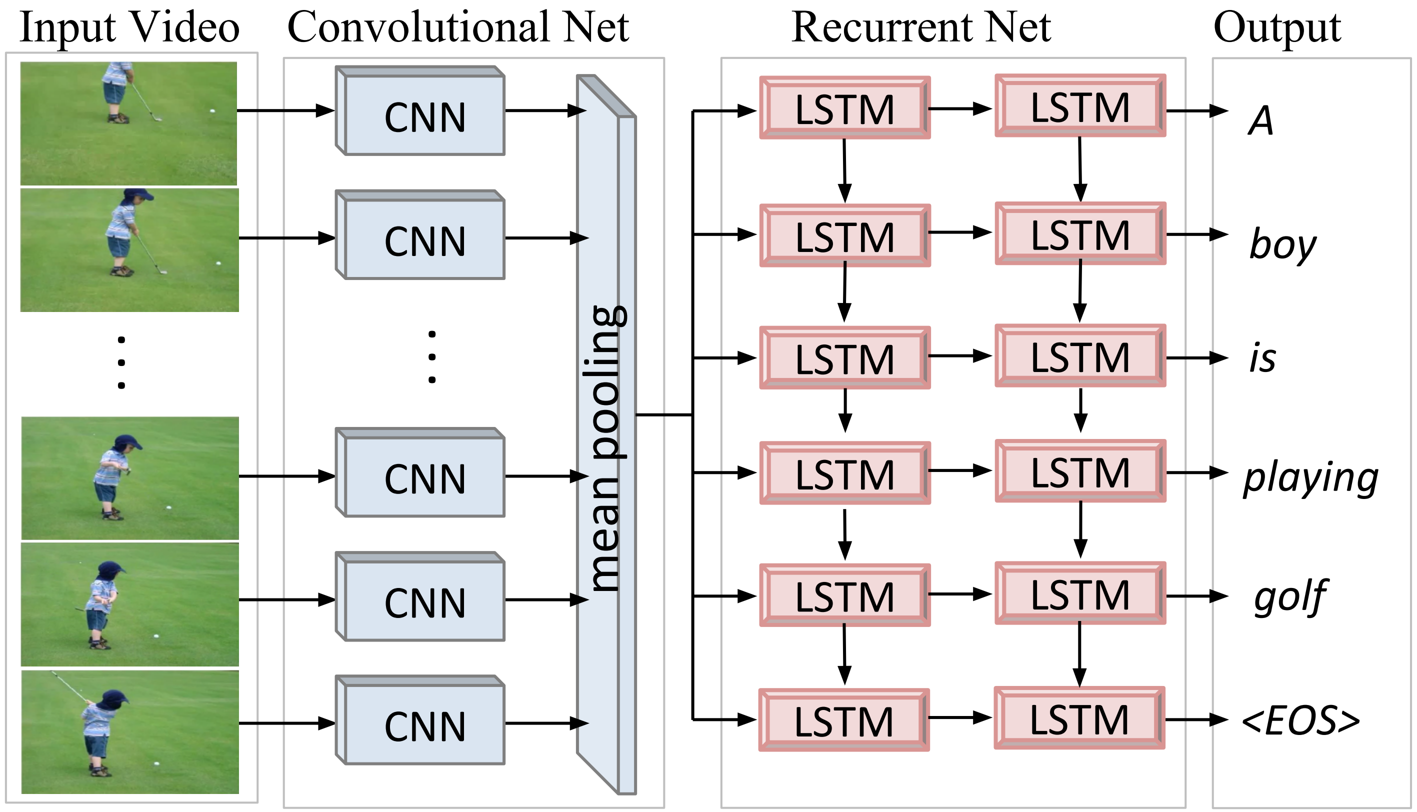
A: No, it's a woman

[Das et al., 2017]

Video Captioning

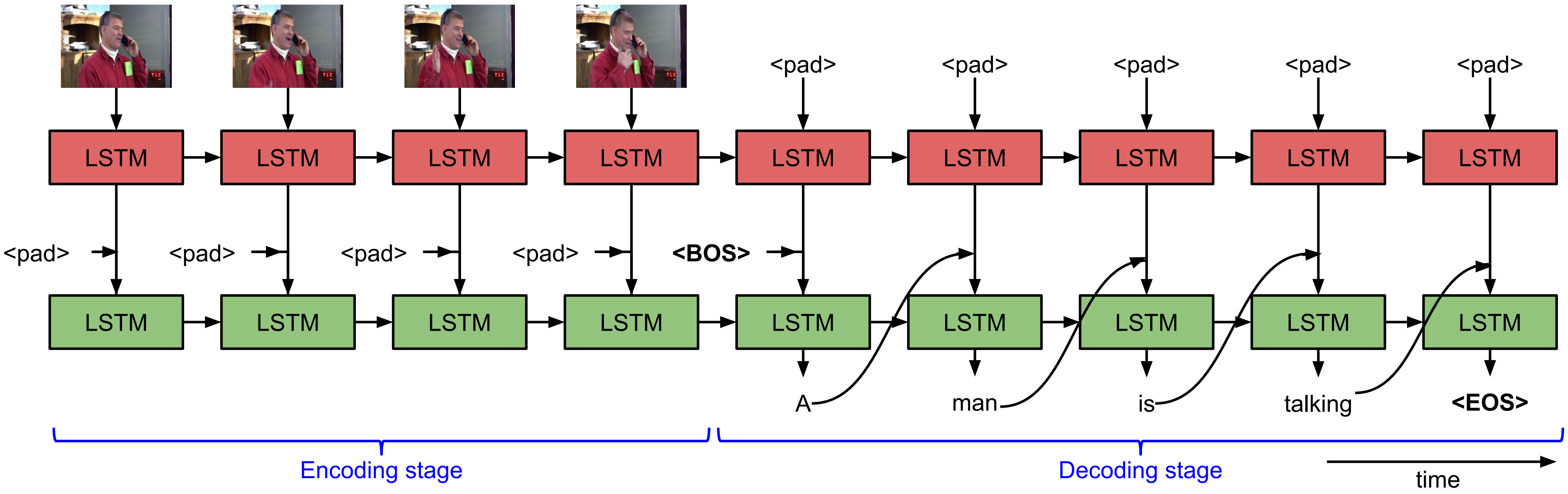
Ground truth: A woman is slicing a red pepper.

Early Video Captioning



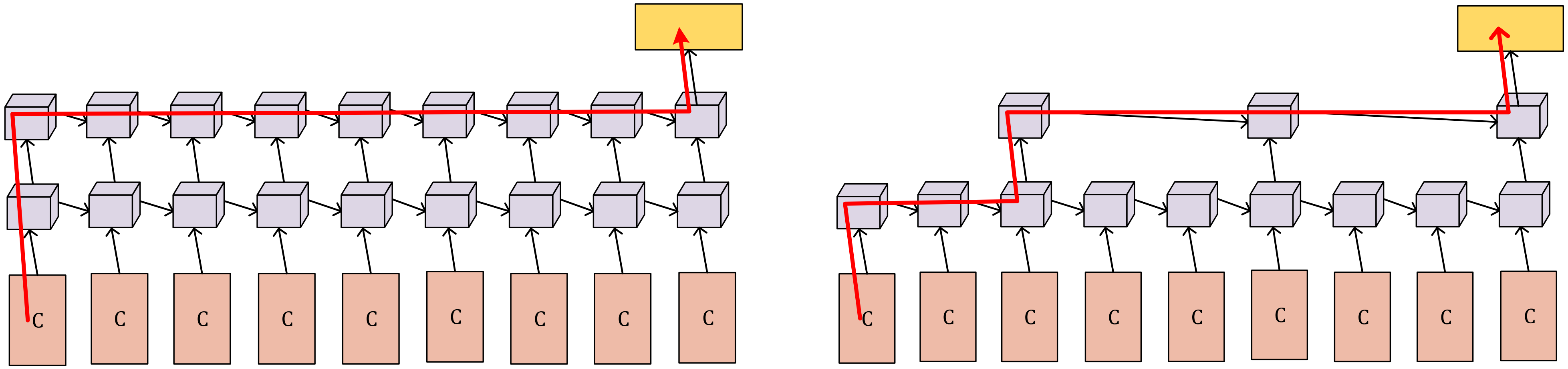
[Venugopalan et al., 2014]

S2VT



[Venugopalan et al., 2015]

Hierarchical Encoder

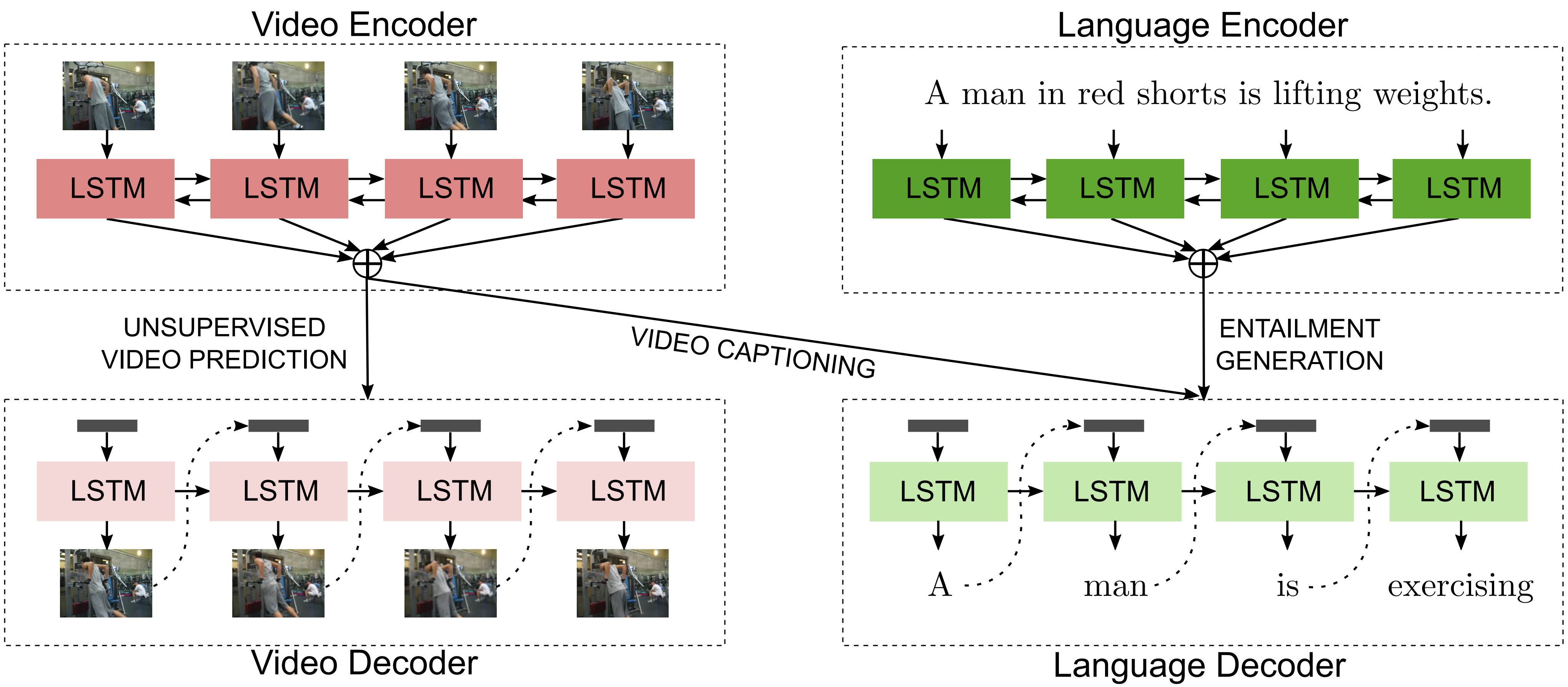


(a) Stacked LSTM video encoder

(b) Hierarchical Recurrent Neural Encoder

[Pan et al., 2016]

M-to-M Multi-Task for Video Captioning



[Pasunuru and Bansal, 2017a]

Guest Research Talk by Ramakanth Pasunuru:

1) Multi-Task Video Captioning with Video and Entailment Generation (ACL 2017)
2) Reinforced Video Captioning with Entailment Rewards (EMNLP 2017)
3) Game-Based Video-Context Dialogue (EMNLP 2018)

Guest Research Talk by Hyounghun Kim:

1) Improving Visual Question Answering by Referring to Generated Paragraph Captions (ACL 2019)
2) Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA (ACL 2020)
3) Modality-Balanced Models for Visual Dialogue (AAAI 2020)

Guest Research Talk by Jie Lei:

1) TVQA: Localized, Compositional Video Question Answering (EMNLP 2018)
2) TVQA+: Spatio-Temporal Grounding for Video Question Answering (ACL 2020)

Guest Research Talk by Darryl Hannan (moved to next week):

1) ManyModalQA: Modality Disambiguation and QA over Diverse Inputs (AAAI 2020)

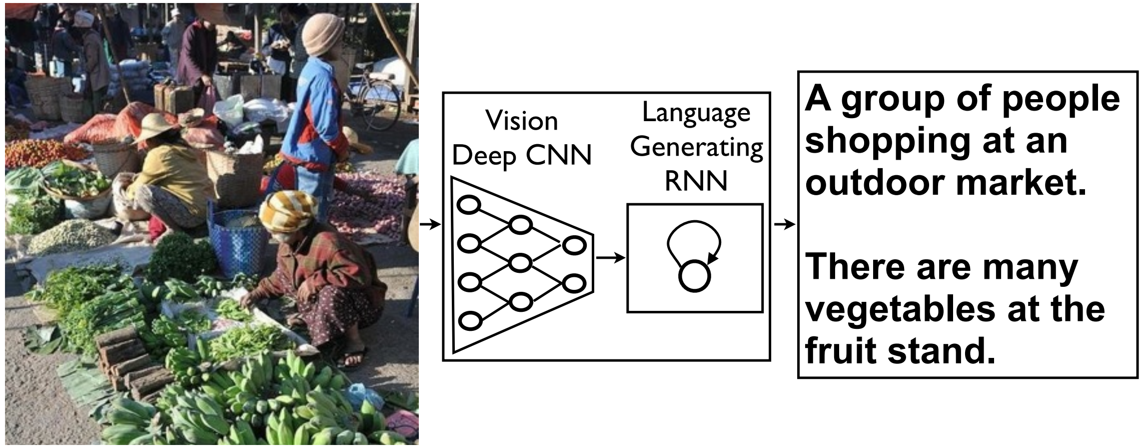
Guest Research Talk by Jaemin Cho (moved to next week):

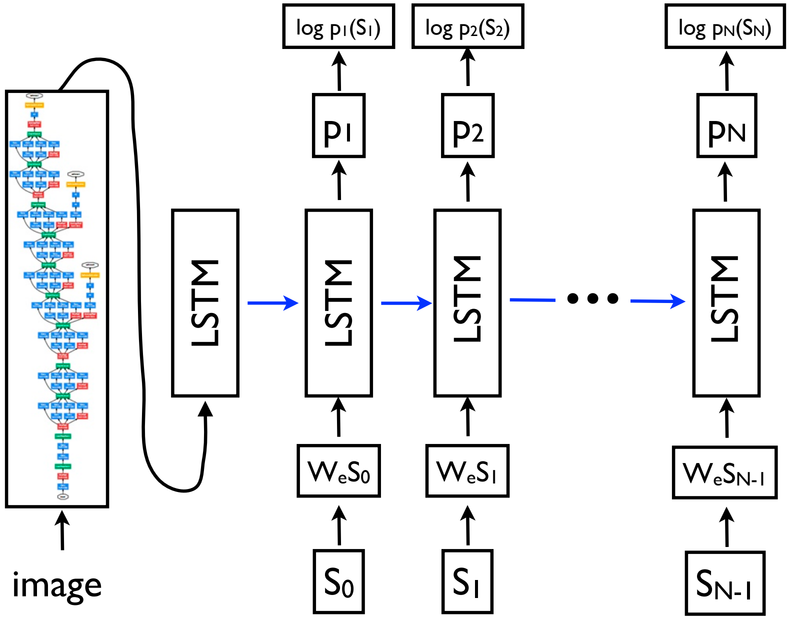
1) LXMERT: Learning Cross-Modality Encoder Representations from Transformers (EMNLP 2019)
2) X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers (EMNLP 2020)

Topic:
Multi-Task and Reinforcement Learning for Video captioning
Game-Based Video-Context Dialogue

(presented by Ramakanth Pasunuru)

Image Captioning

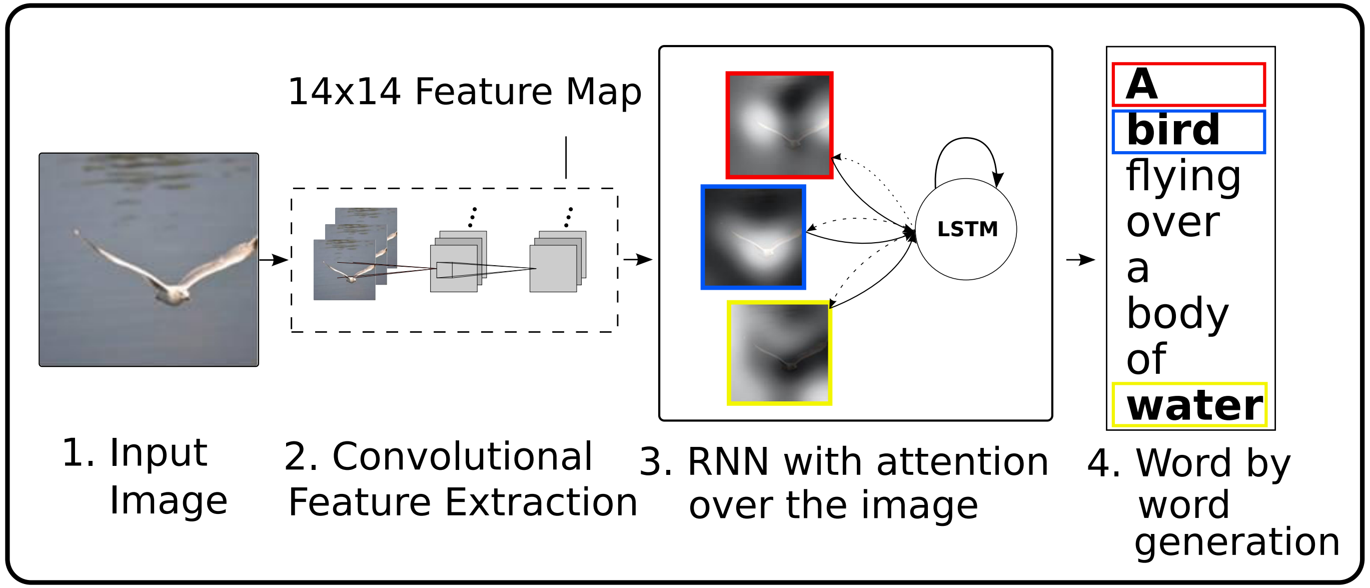




Architecture of Image Captioning Model

[Vinyals et al., 2015]

Image Captioning with Attention



$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.$$

[Xu et al., 2016]

Examples

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

[Xu et al., 2016]

Video Captioning

Ground truth: A woman is slicing a red pepper.

Ground truth: A group of boys are fighting.

Applications:

- Assistance to visually impaired
- Improving online video search
- Grounded robotic instruction tasks

[Kojima et al., 2002; Lee et al., 2008; Khan and Gotoh, 2012; Barbu et al., 2012; Das et al., 2013; Rohrbach et al., 2013; Yu and Siskind, 2013; Venugopalan et al., 2014, 2015, 2016]

Multi-Task for Video Captioning

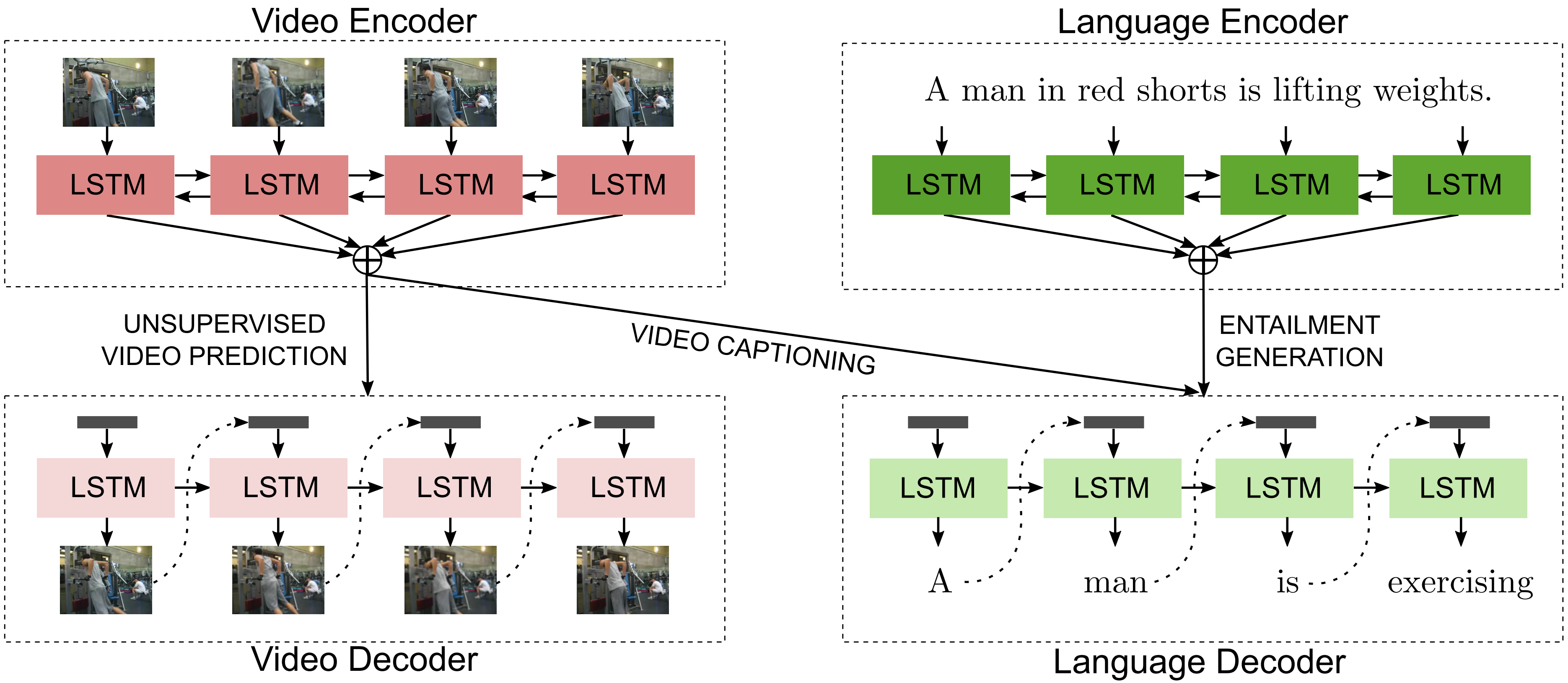
Video Captioning Challenges:
- Lack of sufficient labeled data
- Spatial-visual modeling
- *Logical* storyline dynamics
- *Temporal* across-frame dynamics

Ground truth: A person is mixing powdered ingradients with water.
A woman is mixing flour and water in a bowl.
Our model: A woman is mixing ingredients in a bowl.

We share knowledge w/ 2 related directed-generation tasks/datasets (textual+visual):

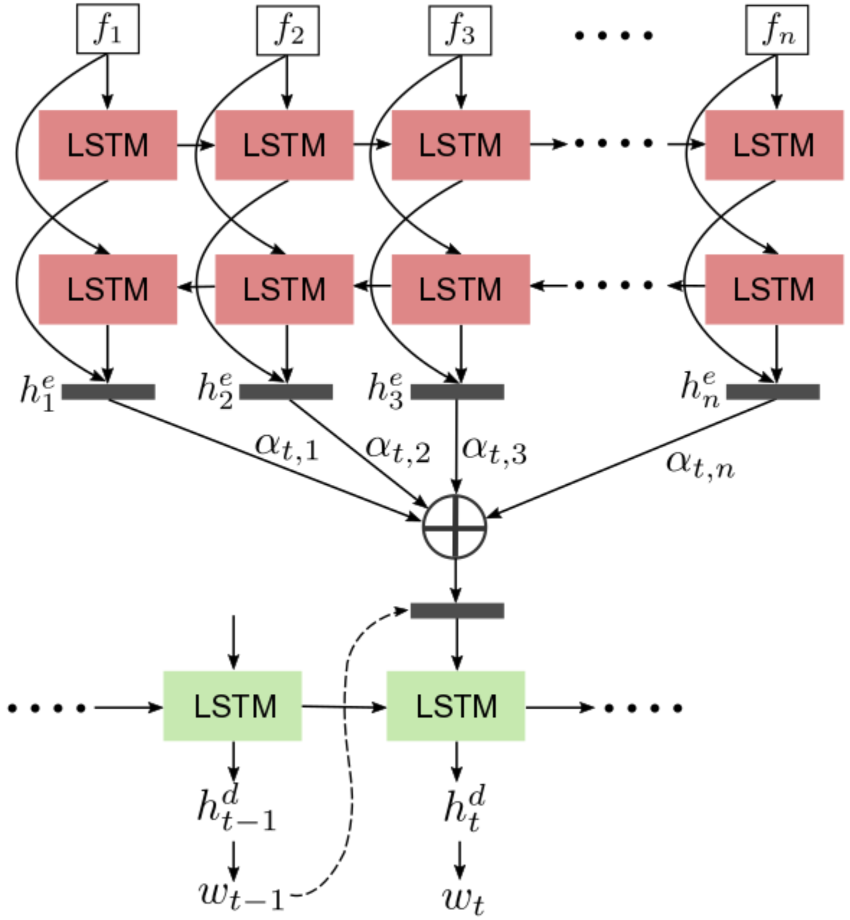
1. Premise-to-Entailment Generation
   - To help learn better caption decoder representations, since caption is also entailed by video.
2. Video-to-Video Generation (Unsupervised)
   - To help learn richer video encoder representations, aware of temporal action context.

Multi-Task for Video Captioning



Baseline Video Captioning Model

- Sequence-to-sequence encoder-decoder model (*'f'* denotes frames; *'w'* denotes words)
- Attention-based (Bahdanau et al., 2015)
- State-of-the-art Inception-v4 image frame features
- Strong baseline (>= previous work)



Textual Entailment

- Directional, logical-implication relation between two sentences:
    - **Premise:** *A girl is jumping on skateboard in the middle of a red bridge.*
    - Entailment: *The girl does a skateboarding trick.*
    - Contradiction: *The girl skates down the sidewalk.*
    - Neutral: *The girl is wearing safety equipment.*

    - **Premise:** *A blond woman is drinking from a public fountain.*
    - Entailment: *The woman is drinking water.*
    - Contradiction: *The woman is drinking coffee.*
    - Neutral: *The woman is very thirsty.*

- Can we use entailment as linguistic inference to help related directed/conditioned generation tasks? (Yes, for e.g. video captioning or document summarization).

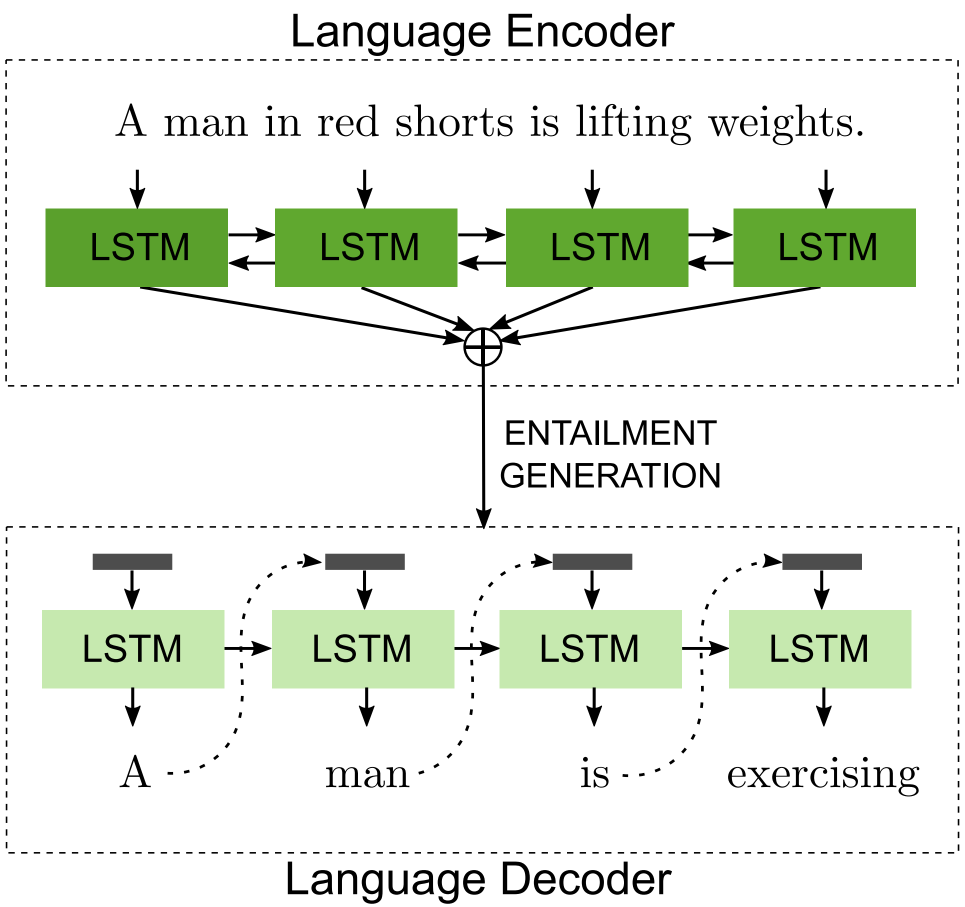
- Large-scale SNLI corpus allows training accurate classification and RNN-style generation models.

[Dagan et al., 2006; Roth and Sammons, 2007; Lai and Hockenmaier, 2014; Bowman et al., 2016]

Entailment Generation

- Helps learn better video-entailing caption decoder representations.

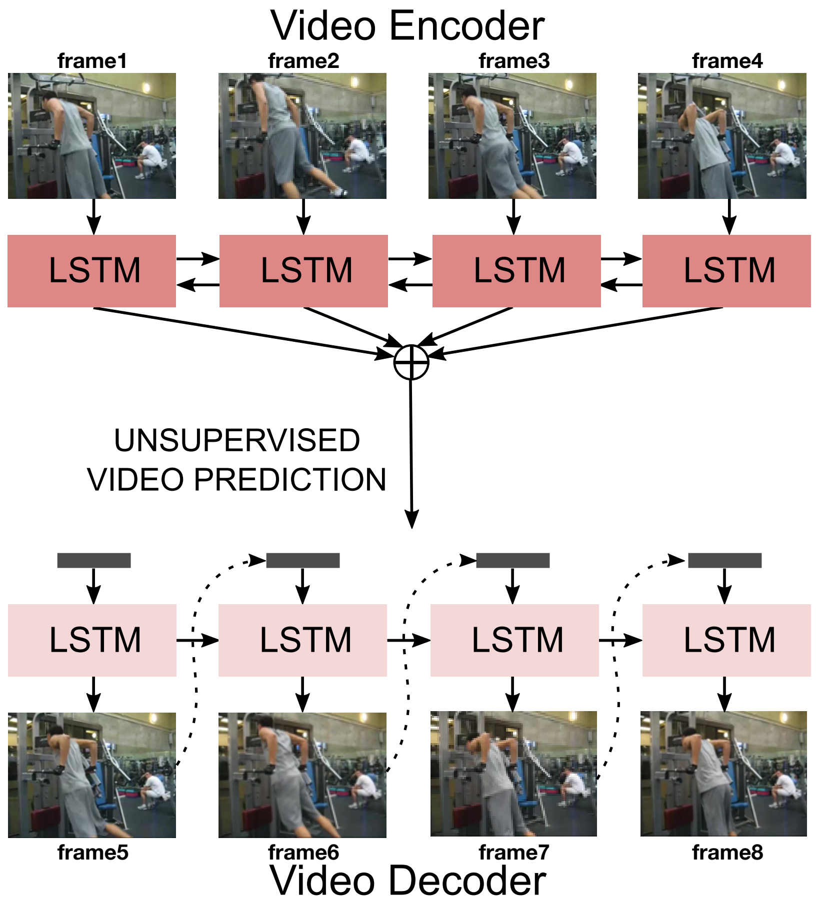
- Since caption needs to be entailed by visual premise of video (i.e., describes subsets of objects/events logically implied by full video content), we teach it about entailment via MTL.



[Kolesnyk et al., 2016]

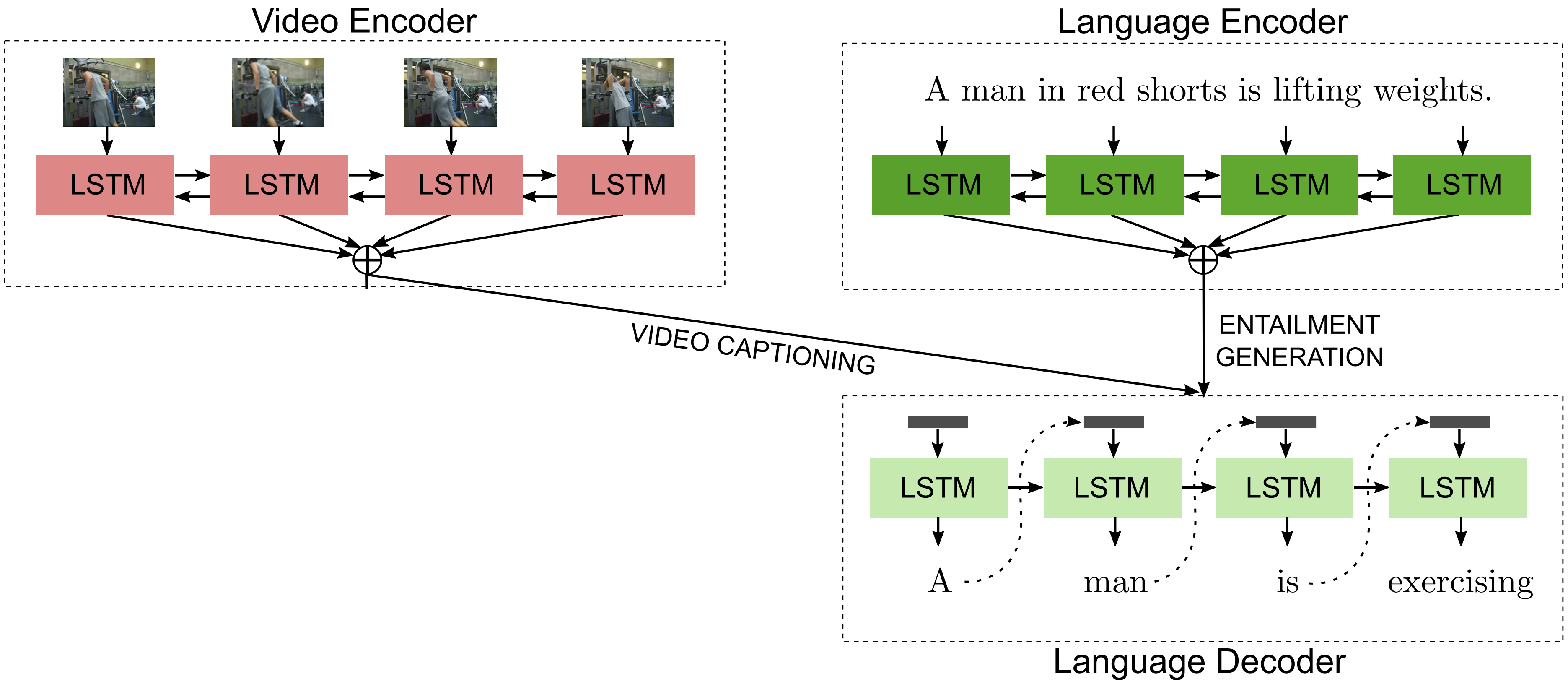
Unsupervised Video Prediction

- Helps learn richer video encoder representations that are aware of temporal context and action sequence/completion.
- Robust to missing frames and varying frame lengths or motion speeds .
- 80:20% frame division between encoder and decoder.
- UCF-101 action videos dataset.

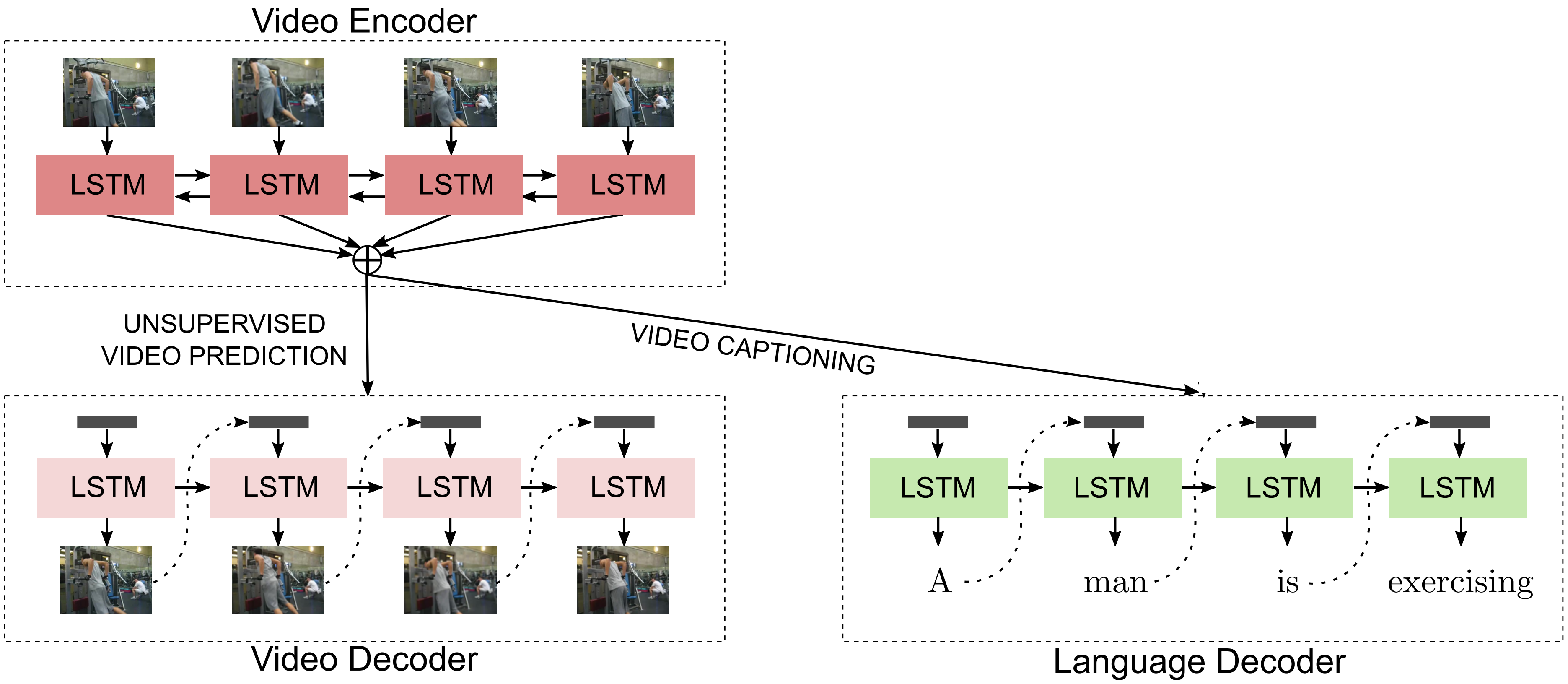


[Srivastava et al., 2015]

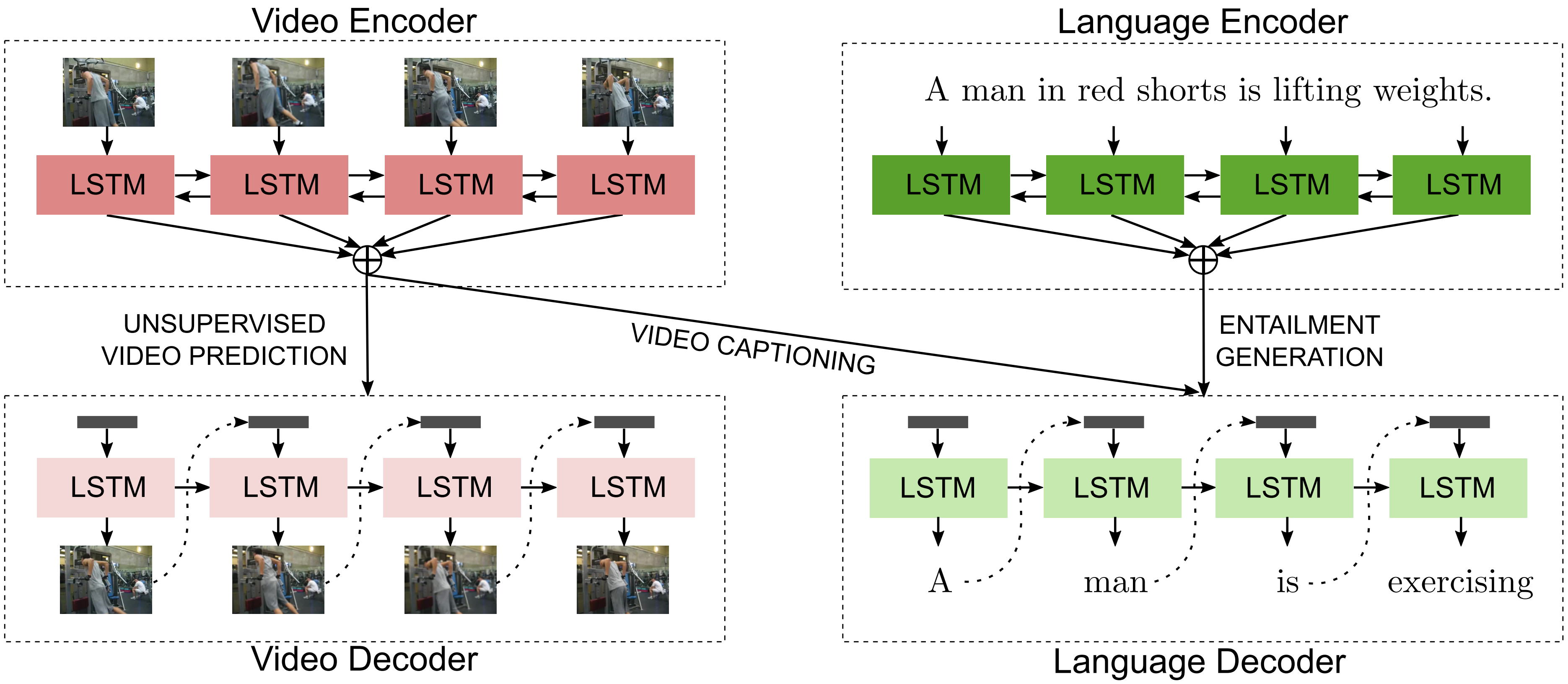
M-to-1 Multi-Task Model



1-to-M Multi-Task Model



M-to-M Multi-Task for Video Captioning



Training in alternate mini-batches: mixing ratio = $\frac{\alpha_v}{(\alpha_v+\alpha_f+\alpha_e)} : \frac{\alpha_f}{(\alpha_v+\alpha_f+\alpha_e)} : \frac{\alpha_e}{(\alpha_v+\alpha_f+\alpha_e)}$

Results (YouTube2Text)

Models	METEOR	CIDEr-D	ROUGE-L	BLEU-4
PREVIOUS WORK				
LSTM-YT (Venugopalan et al., 2015b)	26.9	-	-	31.2
S2VT (Venugopalan et al., 2015a)	29.8	-	-	-
Temporal Attention (Yao et al., 2015)	29.6	51.7	-	41.9
LSTM-E (Pan et al., 2016b)	31.0	-	-	45.3
Glove + DeepFusion (Venugopalan et al., 2016)	31.4	-	-	42.1
p-RNN (Yu et al., 2016)	32.6	65.8	-	49.9
HNRE + Attention (Pan et al., 2016a)	33.9	-	-	46.7
OUR BASELINES				
Baseline (V)	31.4	63.9	68.0	43.6
Baseline (G)	31.7	64.8	68.6	44.1
Baseline (I)	33.3	75.6	69.7	46.3
Baseline + Attention (V)	32.6	72.2	69.0	47.5
Baseline + Attention (G)	33.0	69.4	68.3	44.9
Baseline + Attention (I)	33.8	77.2	70.3	49.9
Baseline + Attention (I) (E) ⊗	35.0	84.4	71.5	52.6

Dataset: 1970 videos with 40 reference captions for each video clip.

Metrics: All the above metrics are automatic based on phrase matching between generated and reference caption. For example, BLEU is based on n-gram matching, ROUGE-L is based on longest common subsequence matching.

Results (YouTube2Text)

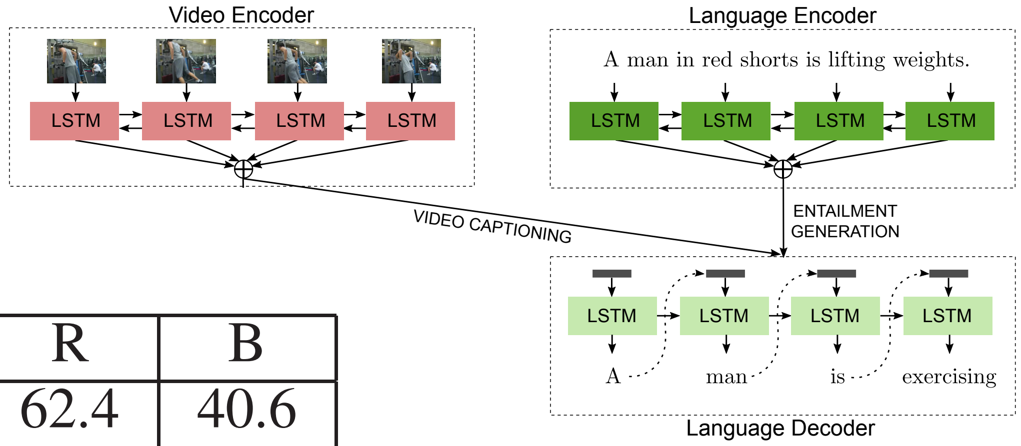
Models	METEOR	CIDEr-D	ROUGE-L	BLEU-4
PREVIOUS WORK				
LSTM-YT (Venugopalan et al., 2015b)	26.9	-	-	31.2
S2VT (Venugopalan et al., 2015a)	29.8	-	-	-
Temporal Attention (Yao et al., 2015)	29.6	51.7	-	41.9
LSTM-E (Pan et al., 2016b)	31.0	-	-	45.3
Glove + DeepFusion (Venugopalan et al., 2016)	31.4	-	-	42.1
p-RNN (Yu et al., 2016)	32.6	65.8	-	49.9
HNRE + Attention (Pan et al., 2016a)	33.9	-	-	46.7
OUR BASELINES				
Baseline (V)	31.4	63.9	68.0	43.6
Baseline (G)	31.7	64.8	68.6	44.1
Baseline (I)	33.3	75.6	69.7	46.3
Baseline + Attention (V)	32.6	72.2	69.0	47.5
Baseline + Attention (G)	33.0	69.4	68.3	44.9
Baseline + Attention (I)	33.8	77.2	70.3	49.9
Baseline + Attention (I) (E) ⊗	35.0	84.4	71.5	52.6
OUR MULTI-TASK LEARNING MODELS				
⊗ + Video Prediction (1-to-M)	35.6	88.1	72.9	54.1
⊗ + Entailment Generation (M-to-1)	35.9	88.0	72.7	54.4
⊗ + Video Prediction + Entailment Gener (M-to-M)	**36.0**	**92.4**	**72.8**	**54.5**

Human evaluation: Multi-task model is better than baseline

* All models (1-to-M, M-to-1 and M-to-M) stat. signif. better than strong SotA baseline.

Results (Entailment Generation)

- Video captioning mutually also helps improve the entailment-generation task in turn (w/ statistical significance).



Models	M	C	R	B
Entailment Generation	29.6	117.8	62.4	40.6
+Video Caption (M-to-1)	**30.0**	**121.6**	**63.9**	**41.6**

Analysis Example

Ground truth: Two men are fighting.
A group of boys are fighting.

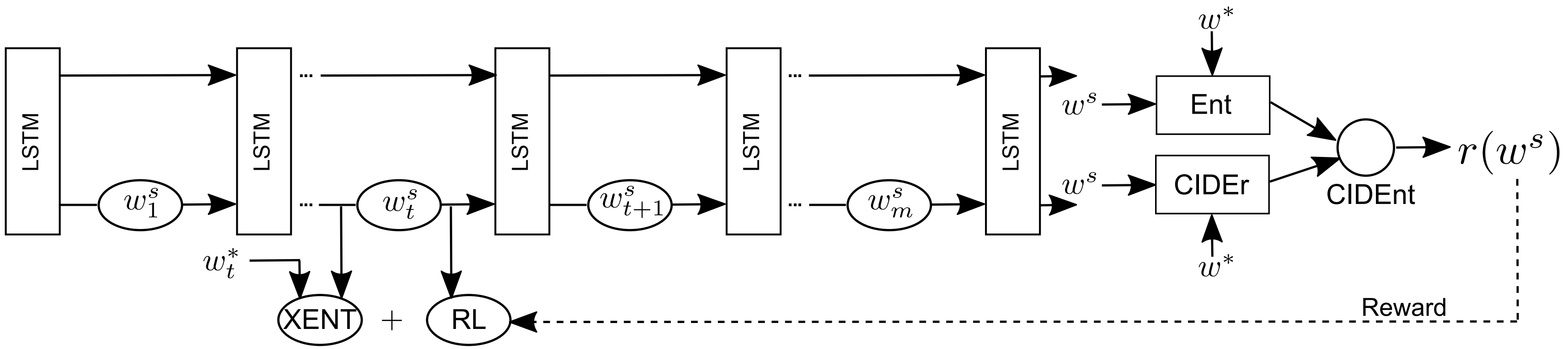
Baseline model: A group of men are dancing.

Multi-task model: Two men are fighting.

Complex example where the multi-task model performs better than baseline.

RL for Video Captioning

We introduce a novel entailment-enhanced reward (CIDEnt) that corrects phrase-matching based metrics (such as CIDEr) to only allow for logically-implied partial matches and avoid contradictions.



[Williams, 1992; Ranzato et al., 2016; Liu et al., 2016; Rennie et al., 2016]

Entailment Corrected Reward

Ground-truth caption	Generated (sampled) caption	CIDEr	Ent
a man is spreading some butter in a pan	puppies is melting butter on the pan	140.5	0.07
a panda is eating some bamboo	a panda is eating some fried	256.8	0.14
a monkey pulls a dogs tail	a monkey pulls a woman	116.4	0.04
a man is cutting the meat	a man is cutting meat into potato	114.3	0.08
the dog is jumping in the snow	a dog is jumping in cucumbers	126.2	0.03
a man and a woman is swimming in the pool	a man and a whale are swimming in a pool	192.5	0.02

Table: Examples of captions sampled during policy gradient and their CIDEr vs Entailment scores.

$$\text{CIDEnt} = \begin{cases} \text{CIDEr} - \lambda, & \text{if} \quad \text{Ent} < \beta \\ \text{CIDEr}, & \text{otherwise} \end{cases}$$

penalizing CIDEr reward when entailment score is low. Thus, ensuring the generated caption logically implies (i.e., is paraphrase or directed partial match w/) ground-truth caption.

Results (MSR-VTT)

Models	BLEU-4	METEOR	ROUGE-L	CIDEr-D	CIDEnt	Human*
PREVIOUS WORK						
Venugopalan (2015b)⋆	32.3	23.4	-	-	-	-
Yao et al. (2015)⋆	35.2	25.2	-	-	-	-
Xu et al. (2016)	36.6	25.9	-	-	-	-
Pasunuru and Bansal (2017)	**40.8**	**28.8**	60.2	47.1	-	-
Rank1: v2t_navigator	**40.8**	28.2	60.9	44.8	-	-
Rank2: Aalto	39.8	26.9	59.8	45.7	-	-
Rank3: VideoLAB	39.1	27.7	60.6	44.1	-	-
OUR MODELS						
Cross-Entropy (Baseline-XE)	38.6	27.7	59.5	44.6	34.4	-
CIDEr-RL	39.1	28.2	60.9	51.0	37.4	11.6
CIDEnt-RL (**New Rank1**)	**40.5**	**28.4**	**61.4**	**51.7**	**44.0**	**18.4**

Table: Our primary video captioning results on MSR-VTT (CIDEnt-RL is stat. significantly better than CIDEr-RL in all metrics, and CIDEr-RL is better than Baseline-XE).

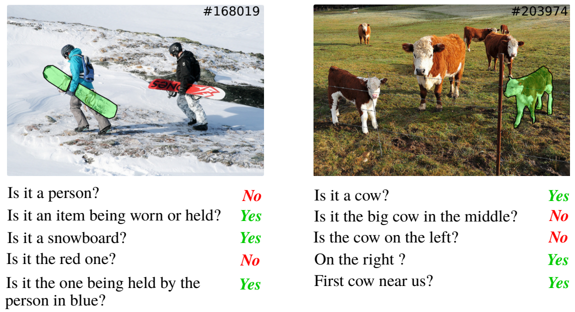
Game-Based Video-Context Dialogue

Visual Context

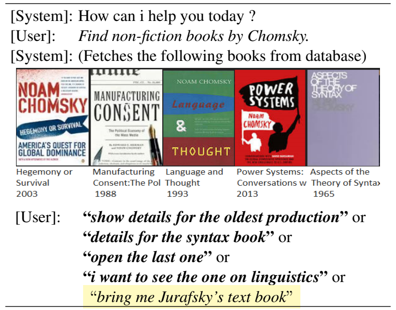
Image-based Context

[Das et al., 2017]

[Mostafazadeh et al., 2017]

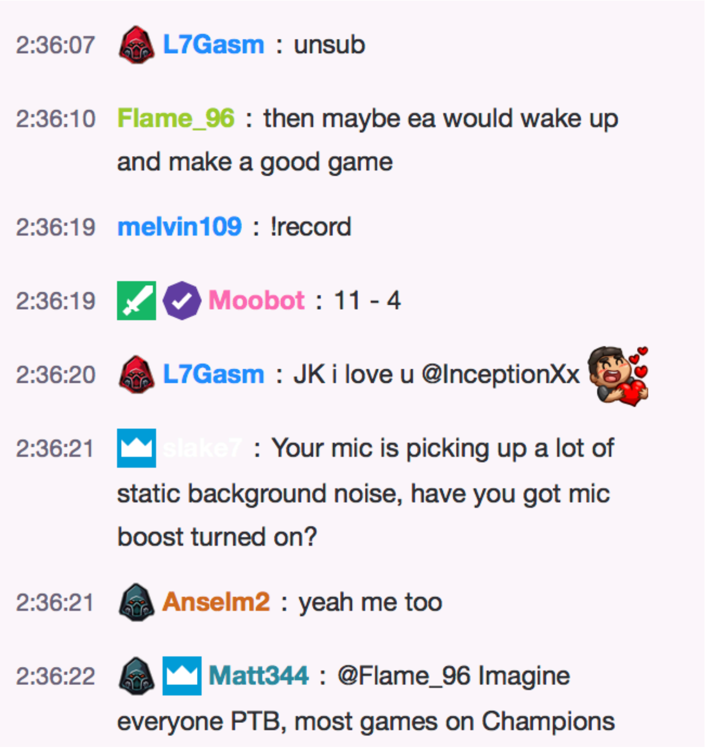


[De Vries et al., 2017]

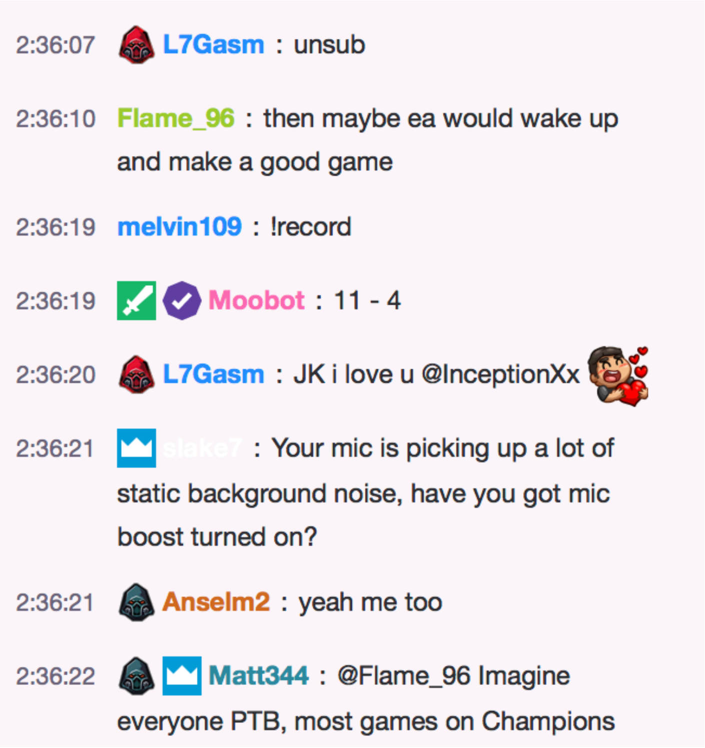


[Celikyilmaz et al., 2014]

Our Twitch-FIFA Dataset


2:36:07 L7Gasm : unsub
2:36:10 Flame_96 : then maybe ea would wake up and make a good game
2:36:19 melvin109 : !record
2:36:19 Moobot : 11 - 4
2:36:20 L7Gasm : JK i love u @InceptionXx
2:36:21 : Your mic is picking up a lot of static background noise, have you got mic boost turned on?
2:36:21 Anselm2 : yeah me too
2:36:22 Matt344 : @Flame_96 Imagine everyone PTB, most games on Champions

Our Twitch-FIFA Dataset



Video + Chat based Context

Multiple speakers

Task

S1: what an offside trap OMEGALUL

S2: Lol that finish bro

S3: suprised you didn't do the extra pass

S4: @S10 a drunk bet? 

S5: @S11 thanks mate

S6: could have passed one more

S7: Pass that

S1: record now!

S8: !record

S9: done a nother pass there

The task is to predict the response (bottom-right) using the video context (left) and the chat context (top-right)

Task

S1: what an offside trap OMEGALUL

S2: Lol that finish bro

S3: suprised you didn't do the extra pass

S4: @S10 a drunk bet? 

S5: @S11 thanks mate

S6: could have passed one more

S7: Pass that

S1: record now!

S8: !record

S9: done a nother pass there

The task is to predict the response (bottom-right) using the video context (left) and the chat context (top-right)

Applications of Video-Grounded Dialogue

- Personal Assistants
- Intelligent tutors
- Human-robot Collaboration

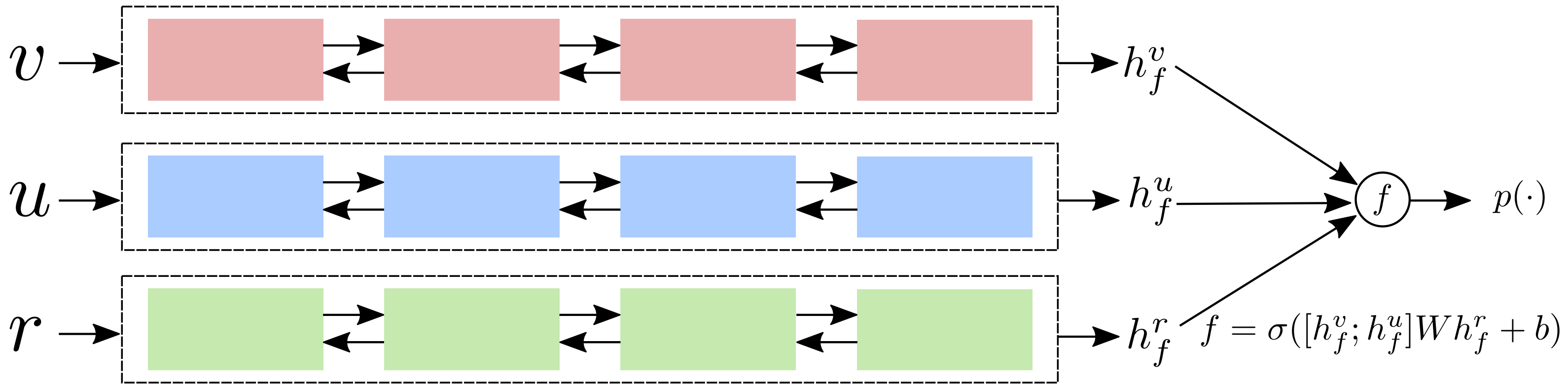
Twitch-FIFA Dataset Statistics

Statistics	Train	Val	Test
#Videos	33	8	8
Total Hours	58.4	11.9	15.4
Final Filtered #Instances	10,510	2,153	2,780
Avg. Chat Context Length	69.0	63.5	71.2
Avg. Response Length	6.5	6.5	6.1

Twitch-FIFA dataset's chat statistics (lengths are defined in terms of number of words)

- Anonymized user identities

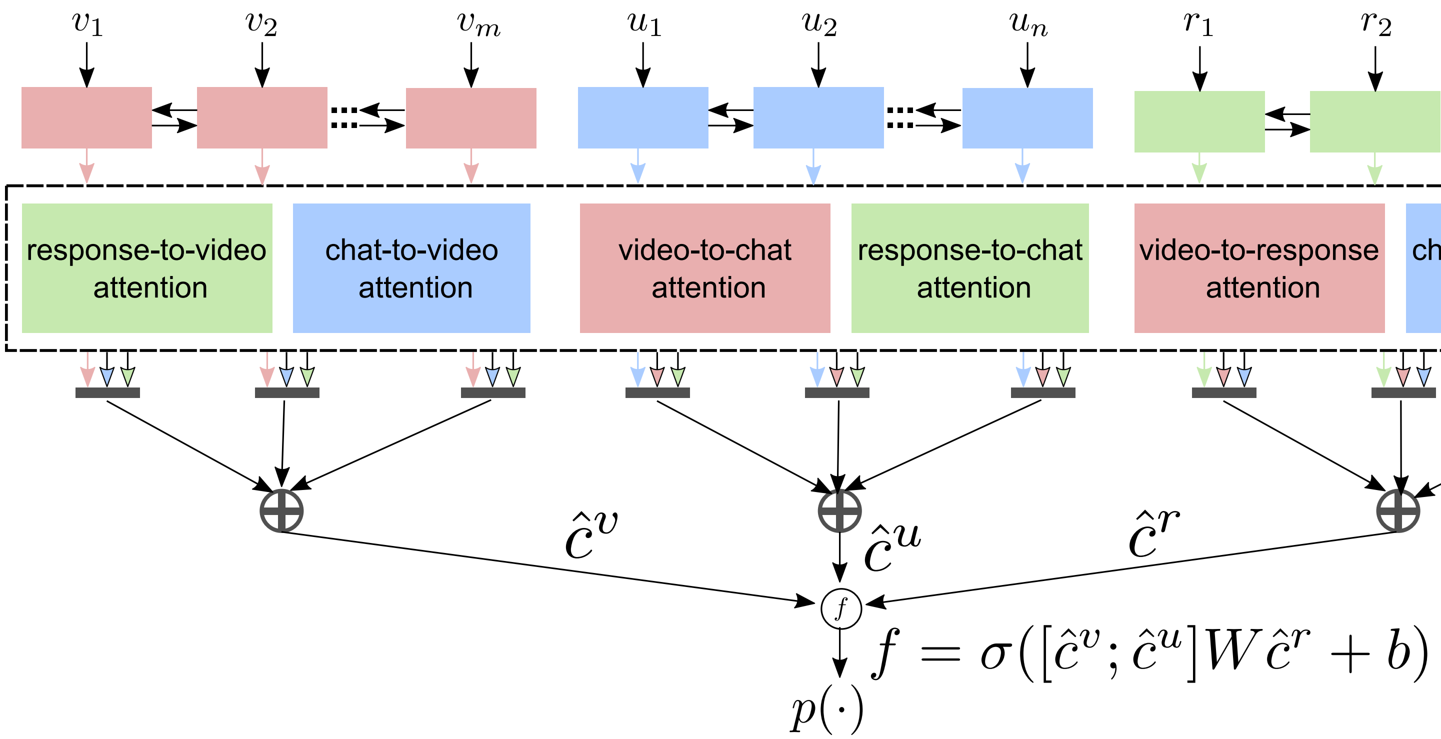
Discriminative Model



Our **Triple Encoder** discriminative model with bidirectional LSTM-RNN encoders for video, chat context, and response

[Lowe et al., 2015]

Discriminative Model



Our **Tri-Directional Attention Flow (TriDAF)** model with all pairwise modality attention modules, as well as self attention on video context, chat context, and response as inputs

[Seo et al., 2017; Lin et al., 2017]

Results

Models	r@1	r@2	r@5
BASELINES			
Most-Frequent-Response	10.0	16.0	20.9
Naive Bayes	9.6	20.9	51.5
Logistic Regression	10.8	21.8	52.5
Nearest Neighbor	11.4	22.6	53.2
Chat-Response-Cosine	11.4	22.0	53.2
DISCRIMINATIVE MODEL			
Dual Encoder (C)	17.1	30.3	61.9
Dual Encoder (V)	16.3	30.5	61.1
Triple Encoder (C+V)	18.1	33.6	68.5
TriDAF+Self Attn (C+V)	20.7	35.3	69.4

Performance of our baselines and discriminative models for recall@k metrics on our Twitch-FIFA test set. C and V represent chat and video context, respectively.

Thank You

(slides by Hyounghun Kim)

Visual QA

Improving Visual Question Answering by Referring to Generated Paragraph Captions

Hyounghun Kim and Mohit Bansal

University of North Carolina at Chapel Hill

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

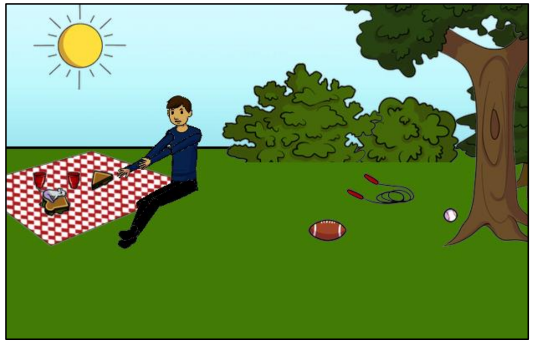
UNC NLP

Main Idea

[1]

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

- Visual question answering is a task to answer diverse questions about images.
- In order to answer all the questions successfully, the ability to understand different aspects of an image is required.

[1] https://visualqa.org/

UNC NLP

Main Idea

[1]

Sentences

1) A girl is eating donuts with a boy in a restaurant
2) A boy and girl sitting at a table with doughnuts.
3) Two kids sitting a coffee shop eating some frosted donuts
4) Two children sitting at a table eating donuts.
5) Two children eat doughnuts at a restaurant table.

Paragraph

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

- Image captioning task is to describe contents or topics from images.
- Singe-sentence captions usually focus on obvious and the most salient part of an image, so tend to describe similar contents.
- On the other hand, paragraph captions contain diverse aspects of an image.

[1]Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. *CVPR, 2017*

(slides by Hyounghun Kim)

Visual Questions vs. Paragraph Captioning (Ground-Truth)

Q. How many planes are in the sky? / A. One
Q. What color are the trees? / A. Green
Q. What color is the plane? / A. White and blue
Q. What color is the sky? / A. Cream and gray
Q. What is in the sky? / A. The plane
Q. What color are the tires? / A. Black
Q. Where was the picture taken? / A. At an airport

GT Paragraph Caption: "The image is of a plane taking off on a runway. There are two planes in the background on the tarmac and one in the sky that has just taken off and is at a very low altitude. The plane that has just taken off is white with blue and gray stripes on it and white writing on the tail. There are trees on the outside of the airport and it is sunset."

Visual Questions vs. Paragraph Captioning (Ground-Truth)

Q. **How many planes are in the sky?** / A. **One**
Q. What color are the trees? / A. Green
Q. What color is the plane? / A. White and blue
Q. What color is the sky? / A. Cream and gray
Q. **What is in the sky?** / A. **The plane**
Q. What color are the tires? / A. Black
Q. **Where was the picture taken?** / A. **At an airport**

GT Paragraph Caption: "The image is of **a plane taking off on a runway.** There are two planes in the background on the tarmac and **one in the sky** that has just taken off and is at a very low altitude. The plane that has just taken off is white with blue and gray stripes on it and white writing on the tail. There are trees on the outside of **the airport** and it is sunset."

Visual Questions vs. Paragraph Captioning (Generated)

Q. How many planes are in the sky? / A. One
Q. What color are the trees? / A. Green
Q. **What color is the plane?** / A. **White and blue**
Q. What color is the sky? / A. Cream and gray
Q. What is in the sky? / A. The plane
Q. What color are the tires? / A. Black
Q. **Where was the picture taken?** / A. **At an airport**

Generated Paragraph Caption: "**A plane is on the runway.** The plane is white. The plane is a plane. The airplane is white. The tail of the plane is red. The sky is very cloudy. The clouds are white. There are trees on the ground. **The planes are white and blue.** The sky is blue."

Manual Investigation (Human Evaluation)

Text	% of Answerable Questions
Ground Truth Caption	55.00
Generated Paragraph Caption	42.67

- Choose random 300 questions.
- Count the questions that can be answered only with text material.

- These results are evidence that a paragraph caption can help VQA task if it can be integrated into VQA model in appropriate ways.

- Paragraph captions provide intermediate textual symbolic evidence for clues.

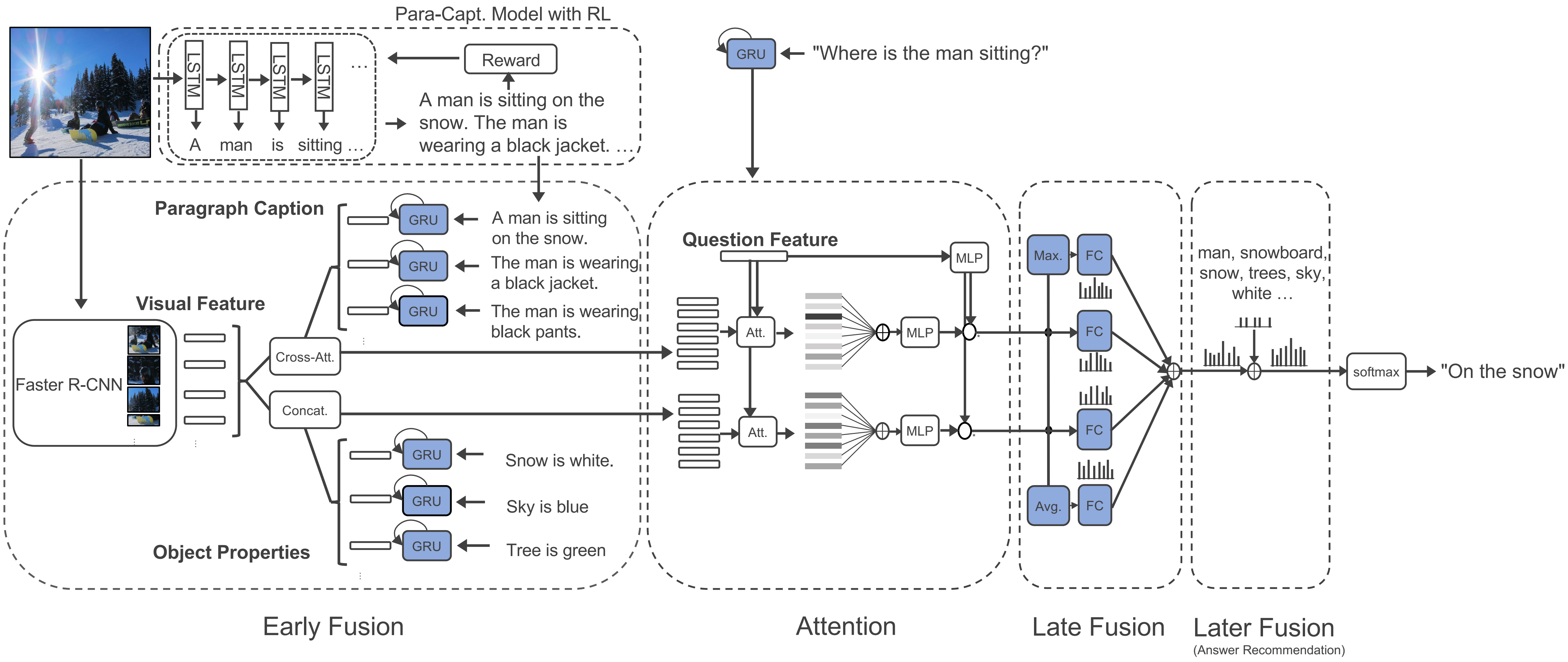
VTQA Model

VTQA (VQA + TextQA)

with Early, Late, and Later Fusion

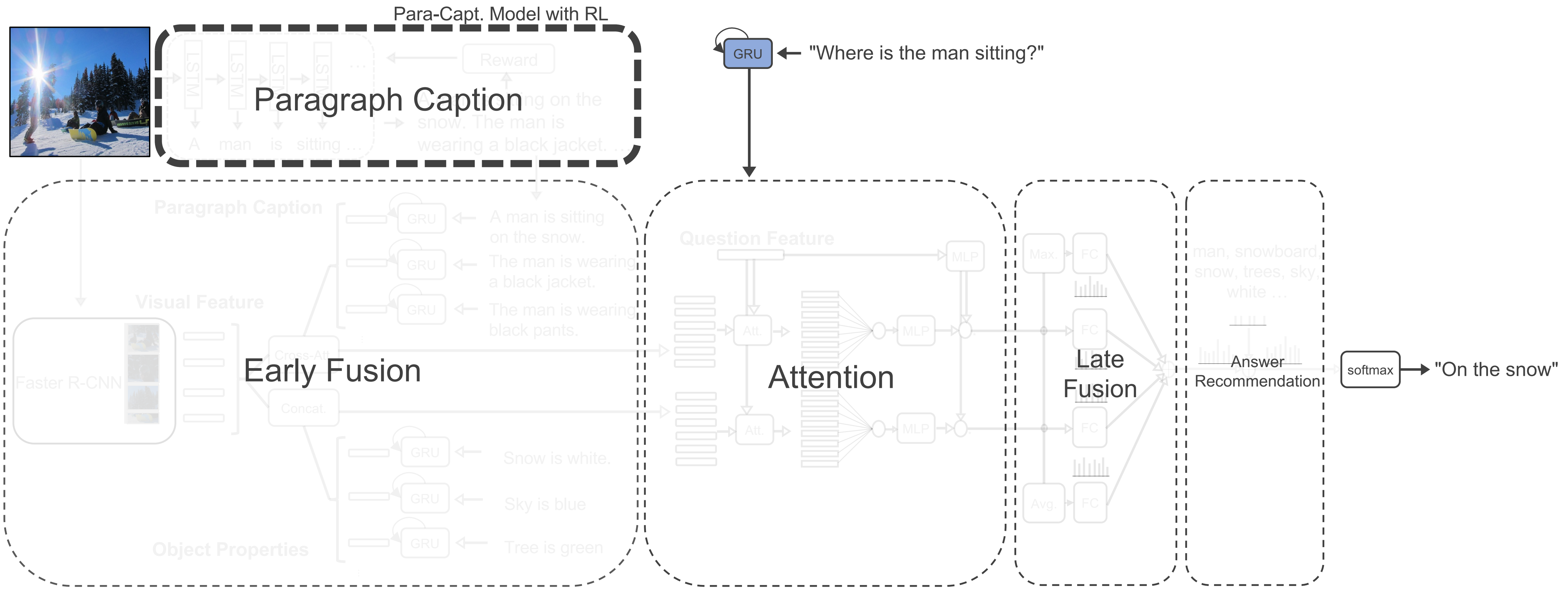
(slides by Hyounghun Kim)

VTQA Model

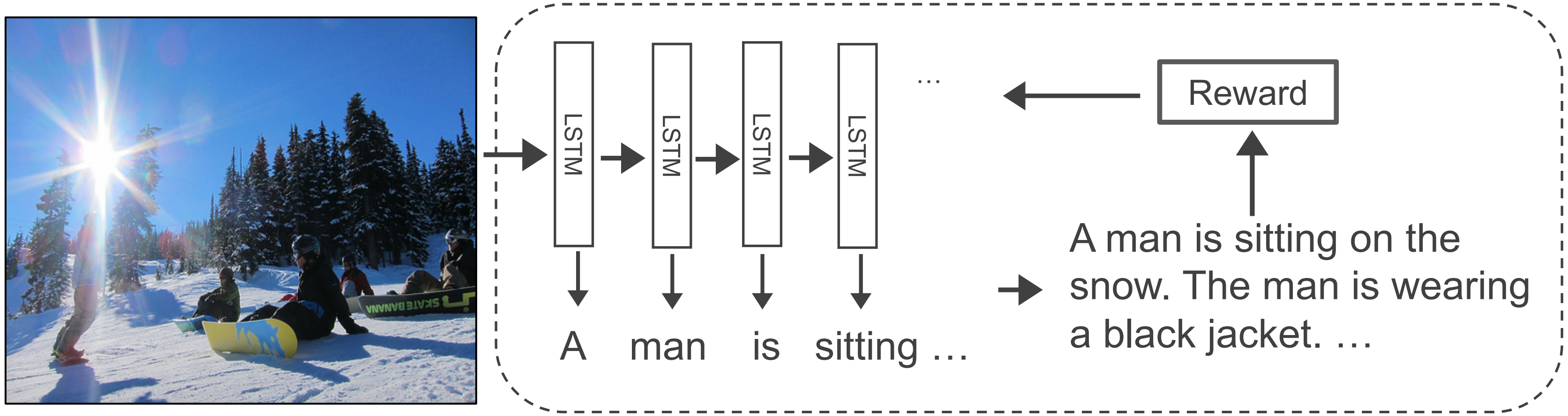


(slides by Hyounghun Kim)

VTQA Model



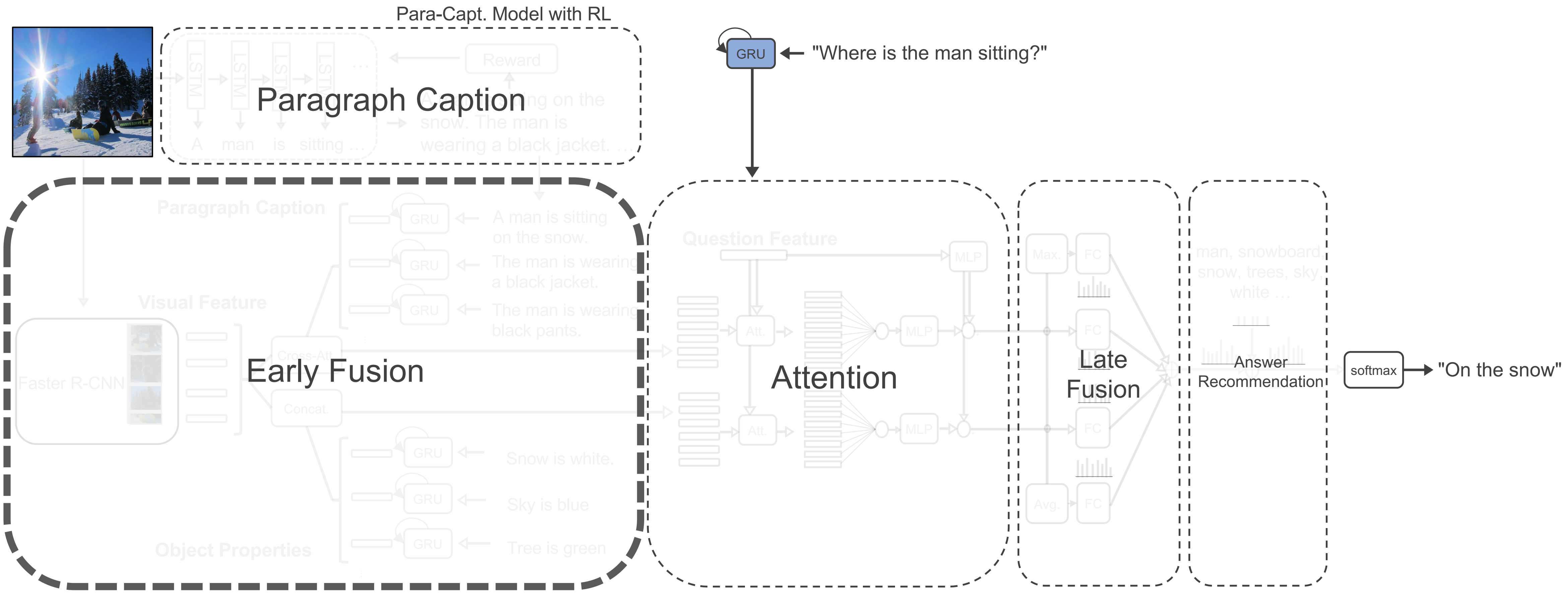
Paragraph Captioning



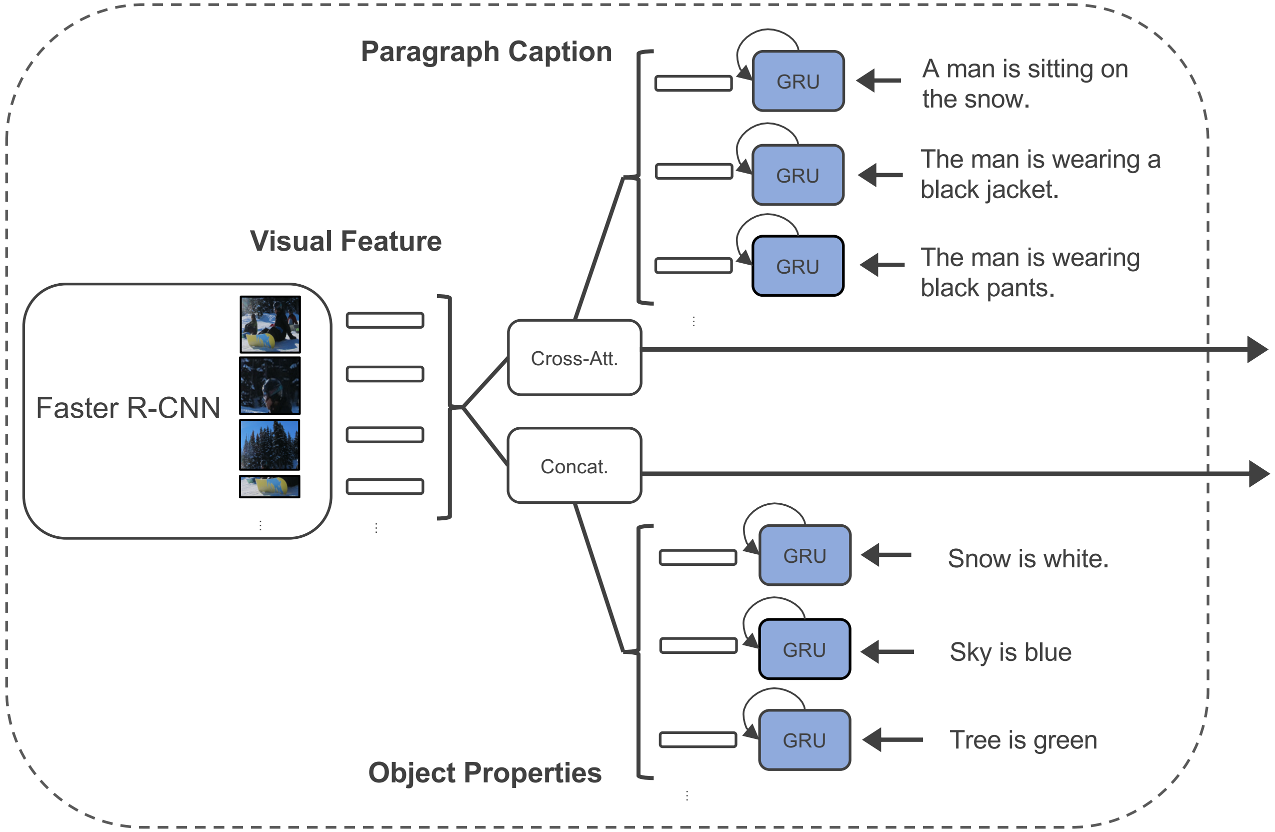
- Paragraph Captioning Model (Melas-Kyriazi et al., 2018).[1]
- Trained with RL using CIDEr-D metric as a reward.
- Repetition penalty applied.
- We will discuss more rewards we tried (saliency, #objects, VQA accuracy) later.

[1]Luke Melas-Kyriazi, Alexander Rush, and George Han. Training for diversity in image paragraph captioning. *EMNLP* 2018.

VTQA Model

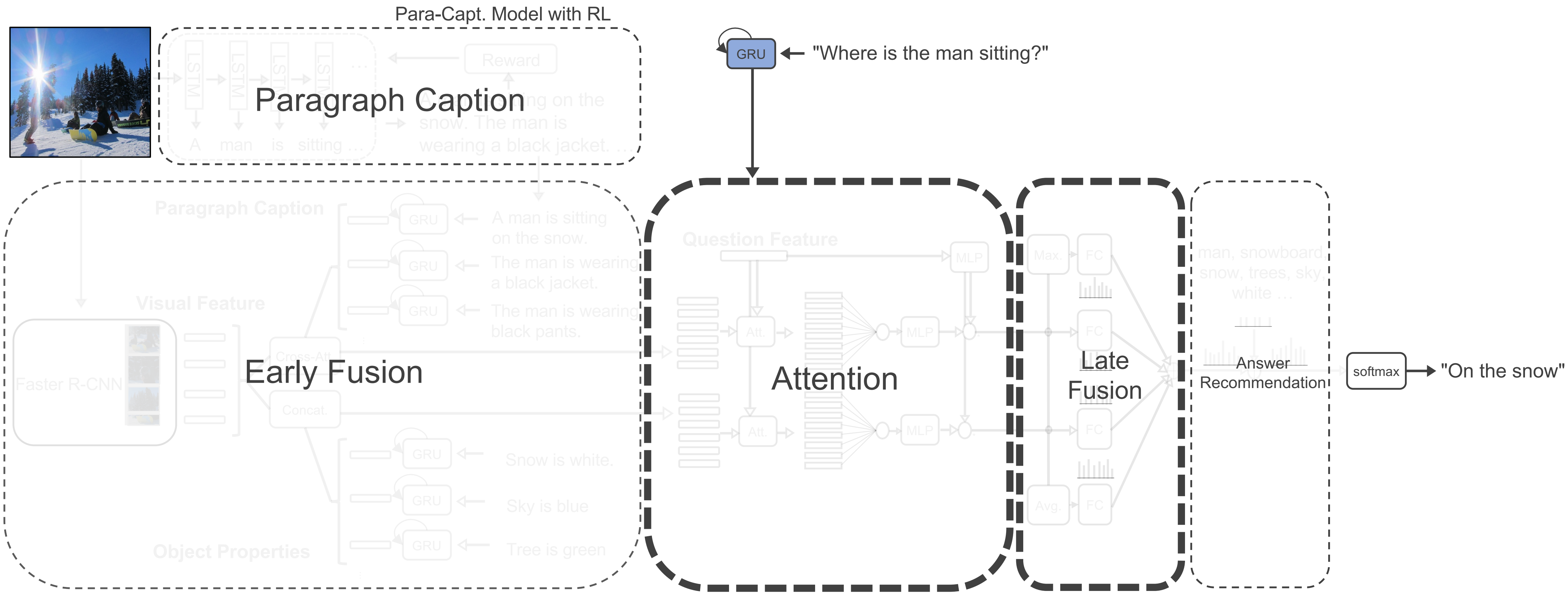

Para-Capt. Model with RL
Paragraph Caption
GRU
"Where is the man sitting?"
Early Fusion
Attention
Late Fusion
Answer Recommendation
softmax
"On the snow"

Early Fusion

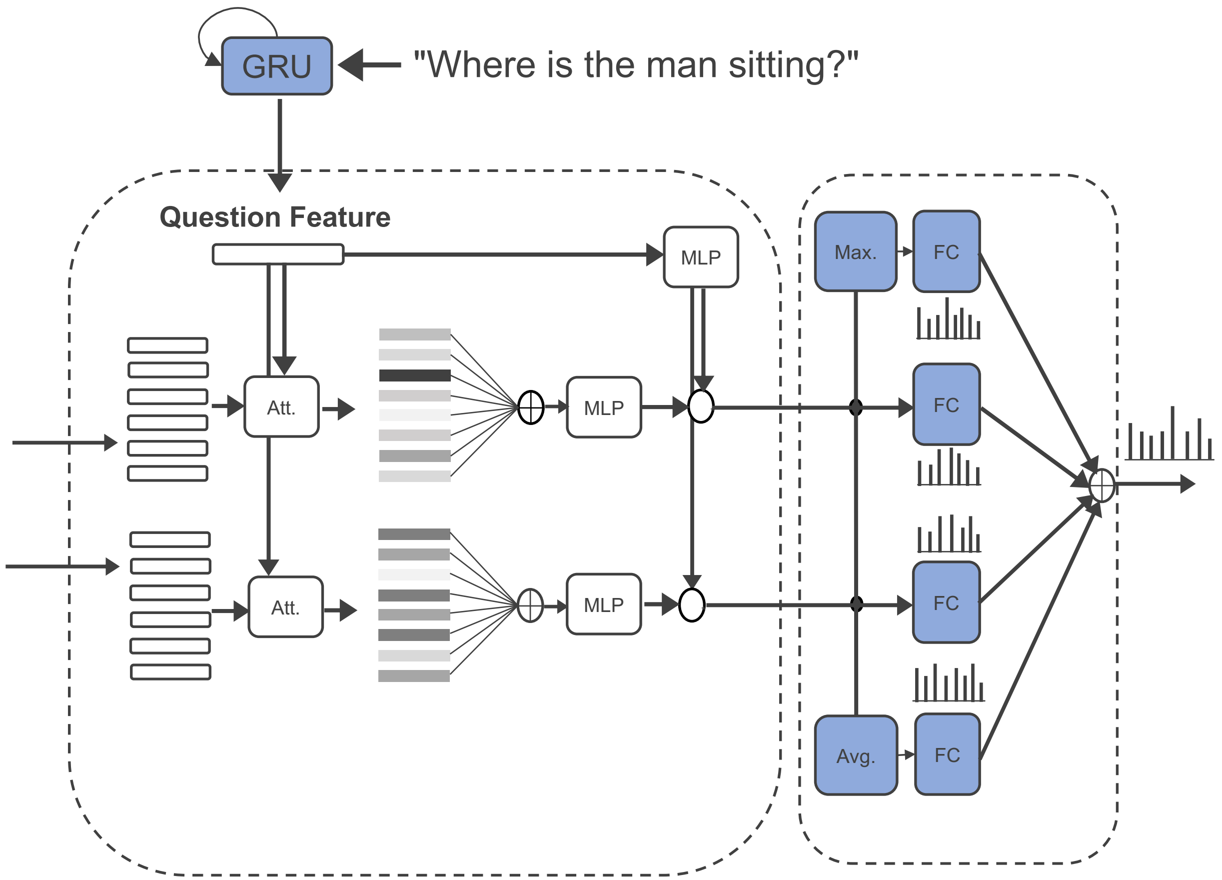


- Faster R-CNN
  - Detects objects in an image.
  - Extracts visual features from each object.
- Cross-Attention
  - Creates a similarity matrix between visual features and paragraph caption features.
  - According to the matrix, relevant features are selected with weights.
- Object Property
  - Encoded with GRU and concatenated to visual features.

VTQA Model

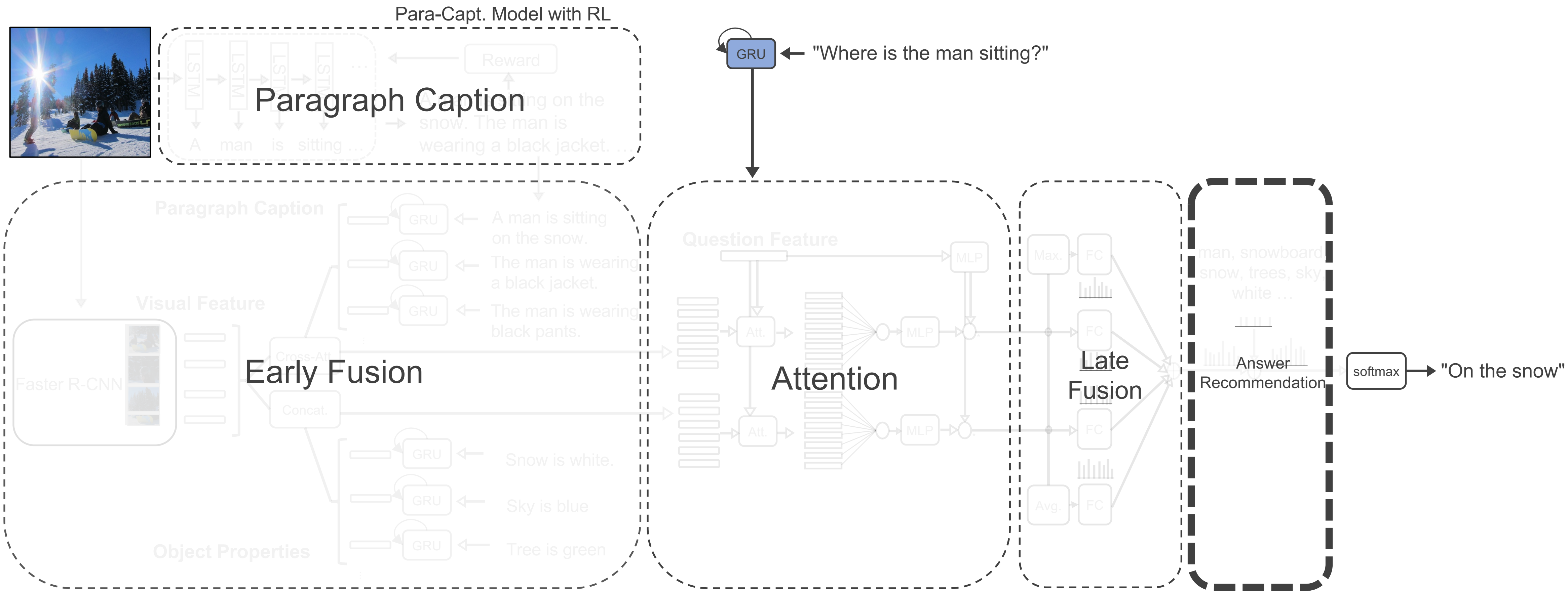

Para-Capt. Model with RL
Paragraph Caption
GRU
"Where is the man sitting?"
Early Fusion
Attention
Late Fusion
Answer Recommendation
softmax
"On the snow"

Attention & Late Fusion

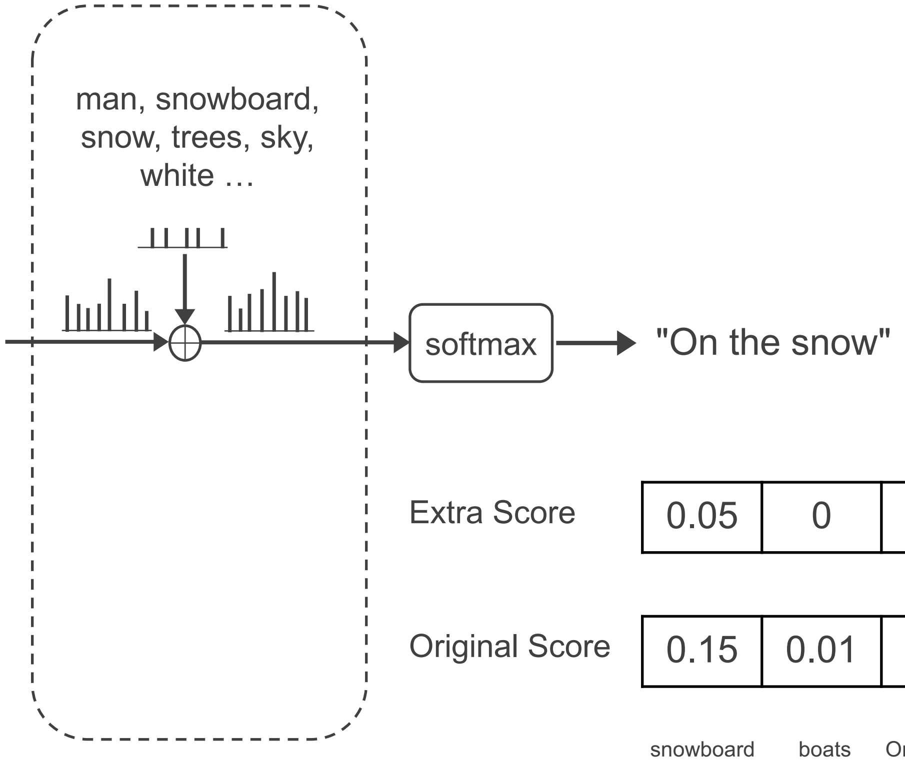


- Attention
  - Question is encoded with GRU.
  - Attention is applied over features from early fusion module w.r.t. a question feature.
- Consensus
  - Each module plays as a voter.
  - The answers that get high scores from multiple voters have a high chance to be selected as the final answer.

VTQA Model


Para-Capt. Model with RL
Paragraph Caption
GRU
"Where is the man sitting?"
Early Fusion
Attention
Late Fusion
Answer Recommendation
softmax
"On the snow"

Later Fusion (Answer Recommendation)



	snowboard	boats	On the snow	water	…
Extra Score	0.05	0	0.05	0	…
+					
Original Score	0.15	0.01	0.21	0.23	…

- Object Property
  - Properties from detected objects can be considered as recommended answers.
  - man, snowboard, white…
- Extra Score
  - An extra score is added to those recommended answers.
  - Extra score = c * standard deviation over all original scores. c is tuned to 1 using validation dataset.

Dataset

- Question-answer pairs from Visual Genome[1]
- Paragraph caption annotations from Krause et al. (2017)[2]
- We follow the image splits of Krause et al. (2017) and exclude those who do not have question-answer pair
- So, the final question-answer pairs split: 171,648 / 29,759 / 29,490 (train / validation / test)

[1]Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia-Li, David Ayman Shamma, Michael Bernstein, Li Fei-Fei

[2]Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. *CVPR, 2017*

UNC NLP

Results

Model	Test Accuracy (%)
VQA baseline	44.68
VQA + MFB baseline	44.94
VTQA (full model)	**46.86**

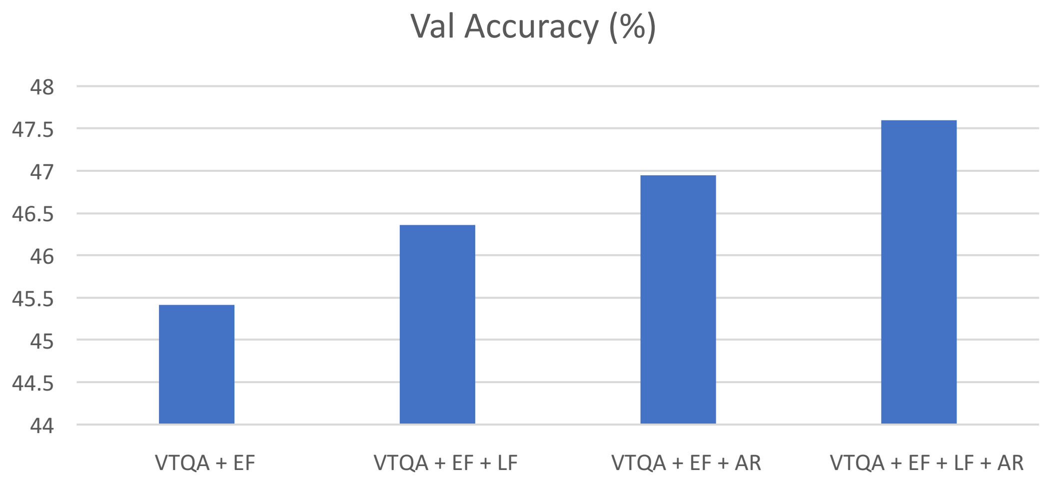
- Run each model 5 times and average them.
- MFB: Multimodal Factorized Bilinear pooling [1]

- Our VTQA model stat. significantly outperforms the baseline VQA model ($p < 0.001$).
- Applied MFB (which is employed in near state-of-the-art models) for comparing with stronger baseline.

[1] Zhou Yu, Jun Yu, Jianping Fan, Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. ICCV 2017.

Ablation Study

Model	Val Accuracy (%)
1. VTQA + EF (base model)	45.41
2. VTQA + EF + LF	46.36
3. VTQA + EF + AR	46.95
4. VTQA + EF + LF + AR	47.60



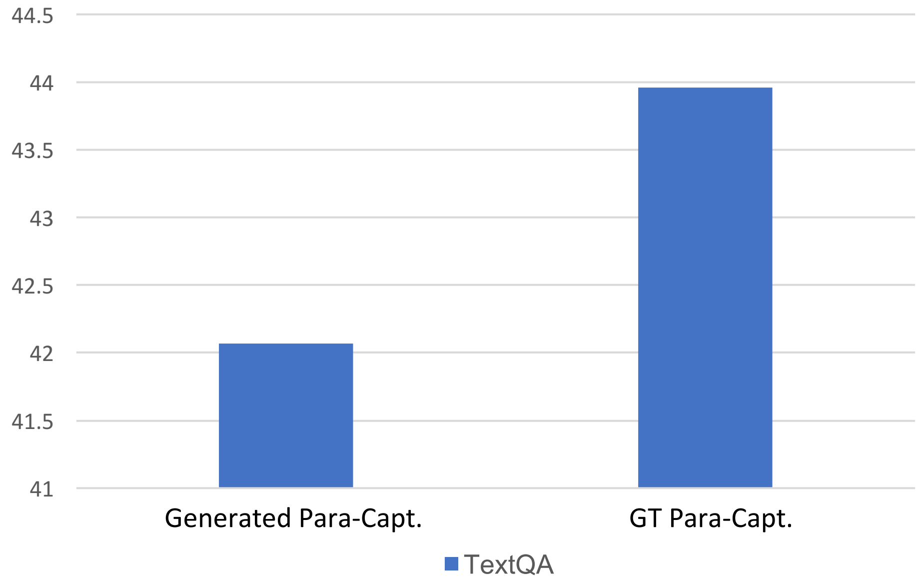
- EF: Early Fusion, LF: Late Fusion, AR: Answer Recommendation.

- Our LF improves the accuracy by 0.95% (from 1 to 2).
- Our AR improves the accuracy by 1.54% (from 1 to 3) and 1.24% (from 2 to 4).

TextQA Model

TextQA	Val Accuracy (%)
GT Para-Capt.	43.96
Generated Para-Capt.	42.07

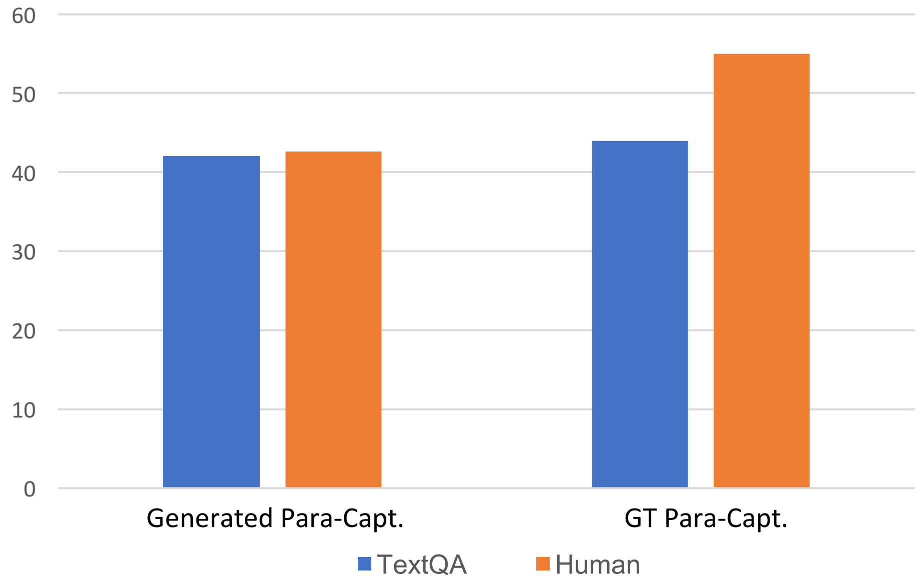
- GT: Ground-Truth



TextQA Model vs. Human Evaluation

TextQA	Val Accuracy (%)
GT Para-Capt.	43.96
Generated Para-Capt.	42.07
Human Eval.	**Accuracy (%)**
GT Para-Capt.	55.00
Generated Para-Capt.	42.67

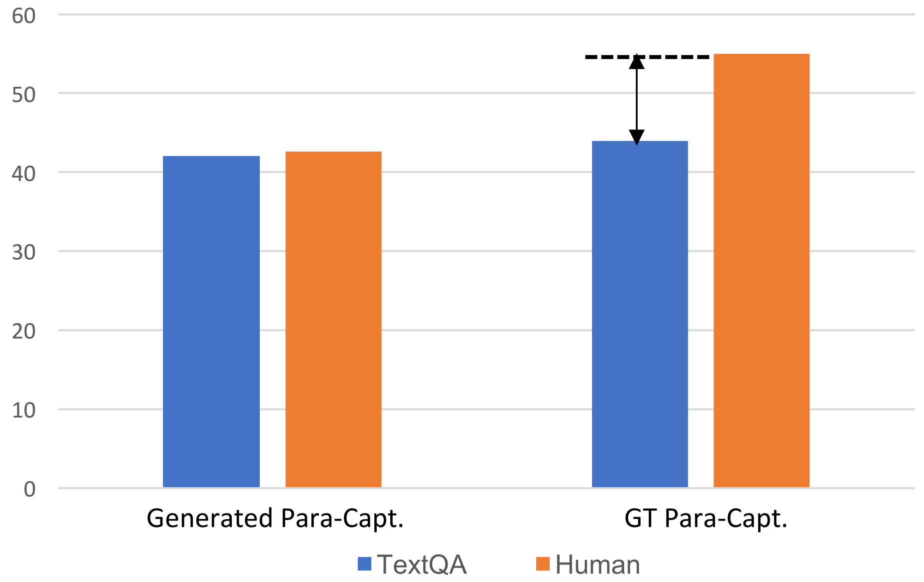
- GT: Ground-Truth



TextQA Model vs. Human Evaluation

TextQA	Val Accuracy (%)
GT Para-Capt.	43.96
Generated Para-Capt.	42.07
Human Eval.	**Accuracy (%)**
GT Para-Capt.	55.00
Generated Para-Capt.	42.67

- GT: Ground-Truth

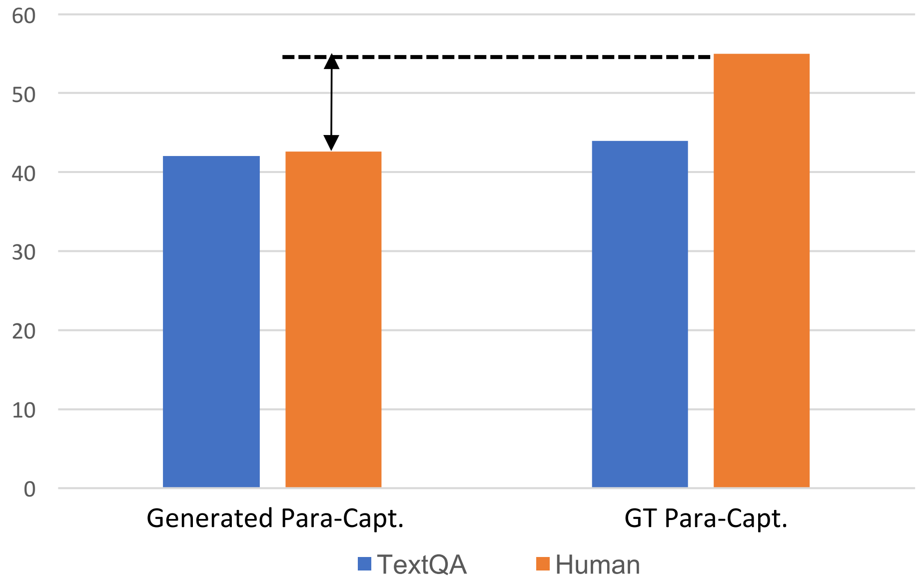


- But, our model does not seem to fully extract information from GT paragraph caption now (43.96 vs. 55.00).

TextQA Model vs. Human Evaluation

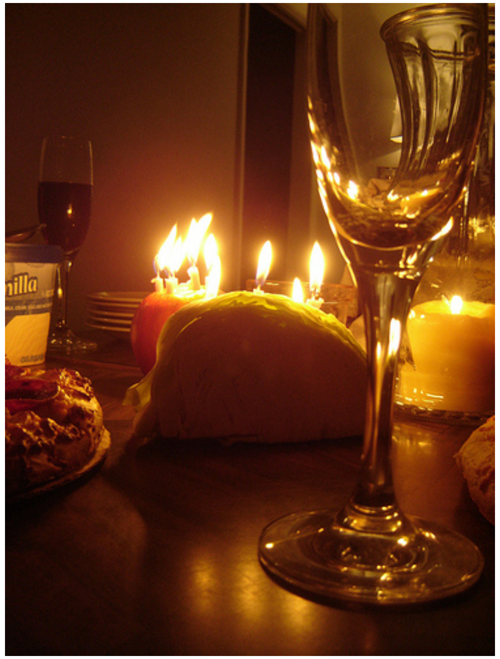
TextQA	Val Accuracy (%)
GT Para-Capt.	43.96
Generated Para-Capt.	42.07
Human Eval.	**Accuracy (%)**
GT Para-Capt.	55.00
Generated Para-Capt.	42.67

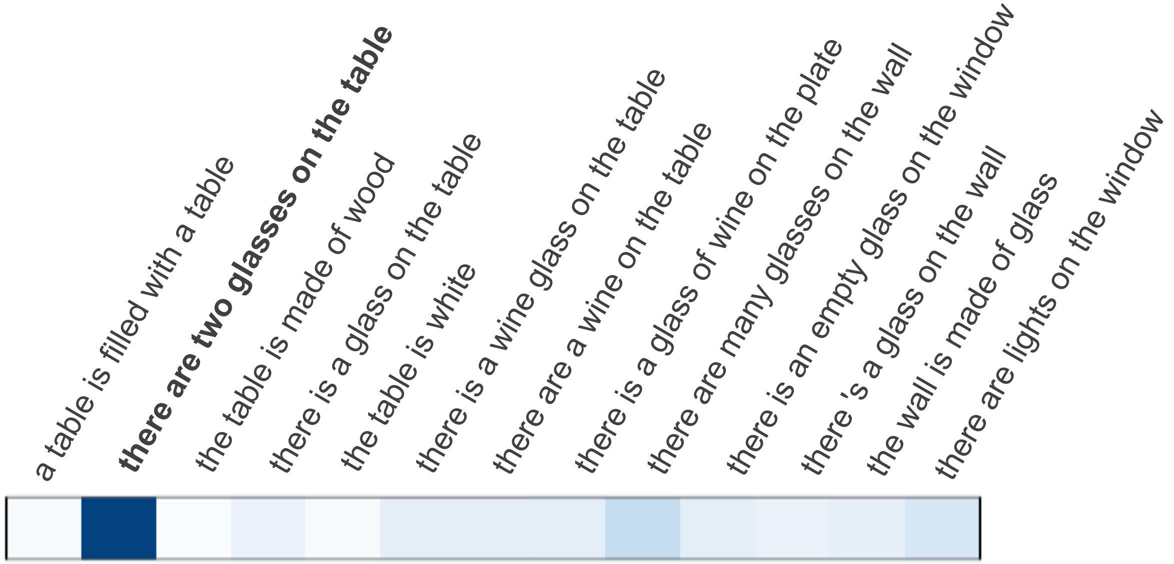
• GT: Ground-Truth



- But, our model does not seem to fully extract information from GT paragraph caption now (43.96 vs. 55.00).
- Also, generated paragraph captions are not good enough to give useful information compared to GT paragraph captions.

Attention Visualization





Q: how many glasses are in the picture A: 2

Q: how many glasses are in the picture A: 2

The paragraph contains a direct clue for the question
→ “there are **two glasses** on the table”

- Examples where image-only VQA model is wrong but our image+para-capt. VTQA model fixes the answer

(slides by Hyounghun Kim)

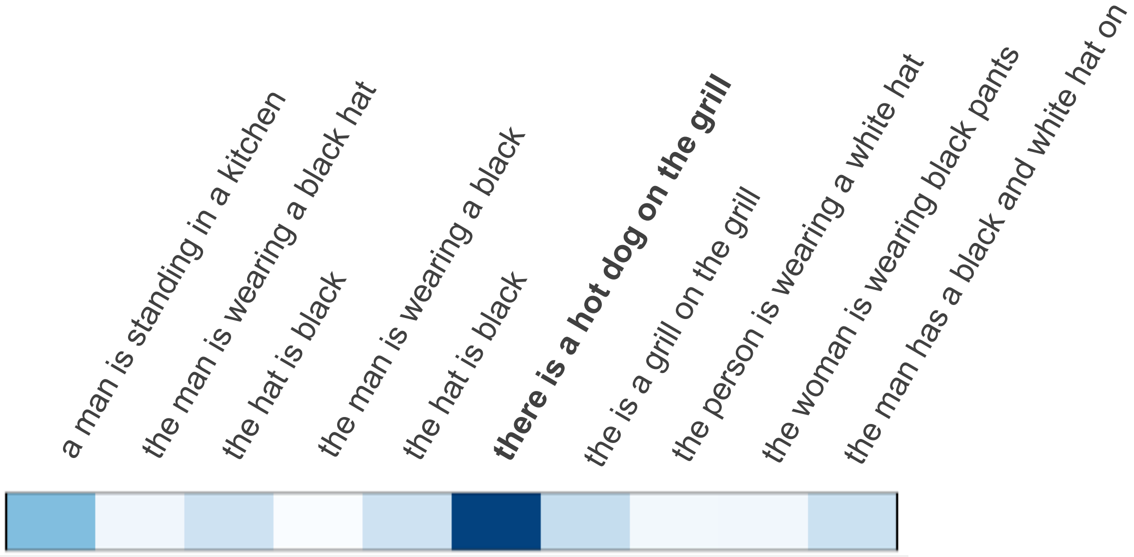
Attention Visualization

Q: where was the photo taken   A: in kitchen

The paragraph contains a direct clue for the question
→ "a young girl is standing **in the kitchen**"

- Examples where image-only VQA model is wrong but our image+para-capt. VTQA model fixes the answer

Attention Visualization



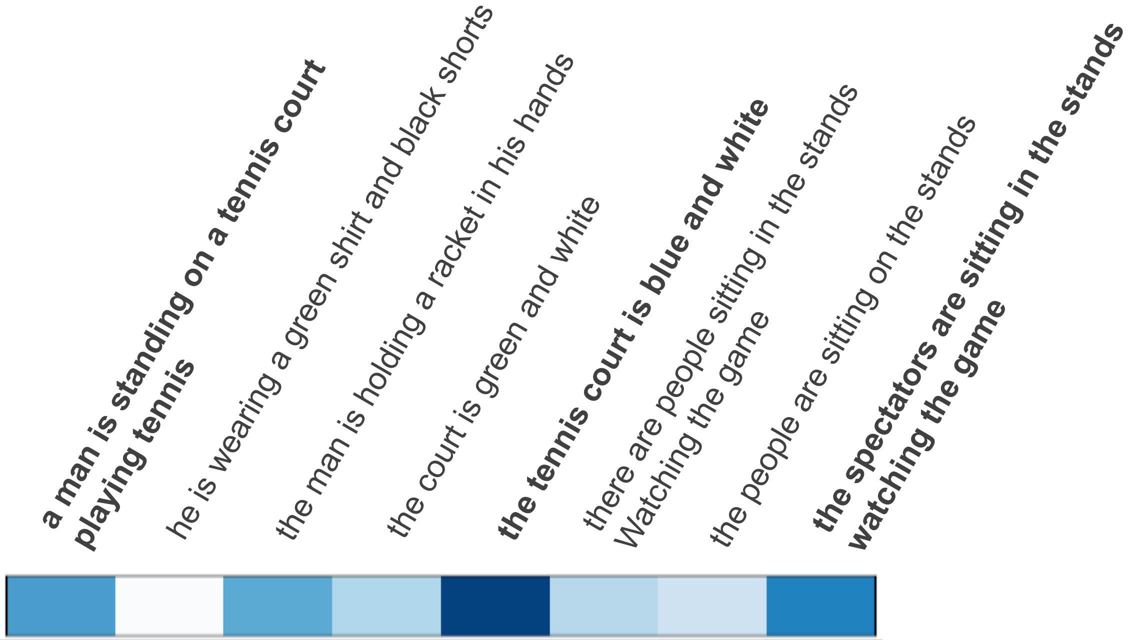
Q: what is being cooked A: hot dogs

Q: what is being cooked A: hot dogs

The paragraph contains a clue that help infer the answer
→ "there is a **hot dog on the grill**"

- Examples where image-only VQA model is wrong but our image+para-capt. VTQA model fixes the answer

Attention Visualization



Q: what is the crowd watching A: tennis match

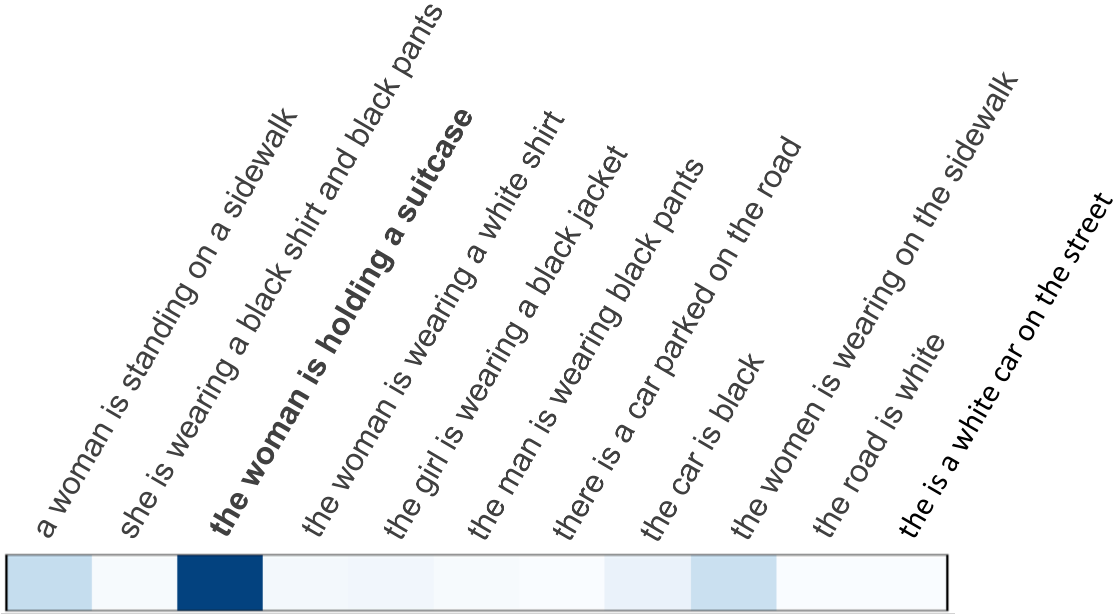
Q: what is the crowd watching A: tennis match

The paragraph has a couple of sentences that give indirect clues
→ "a man is standing on a tennis court **playing tennis**"
→ "the **tennis court** is blue and white"
→ "the spectators are sitting in the stands **watching the game**"

(slides by Hyounghun Kim)

UNC NLP

Attention Visualization (Failed Case)

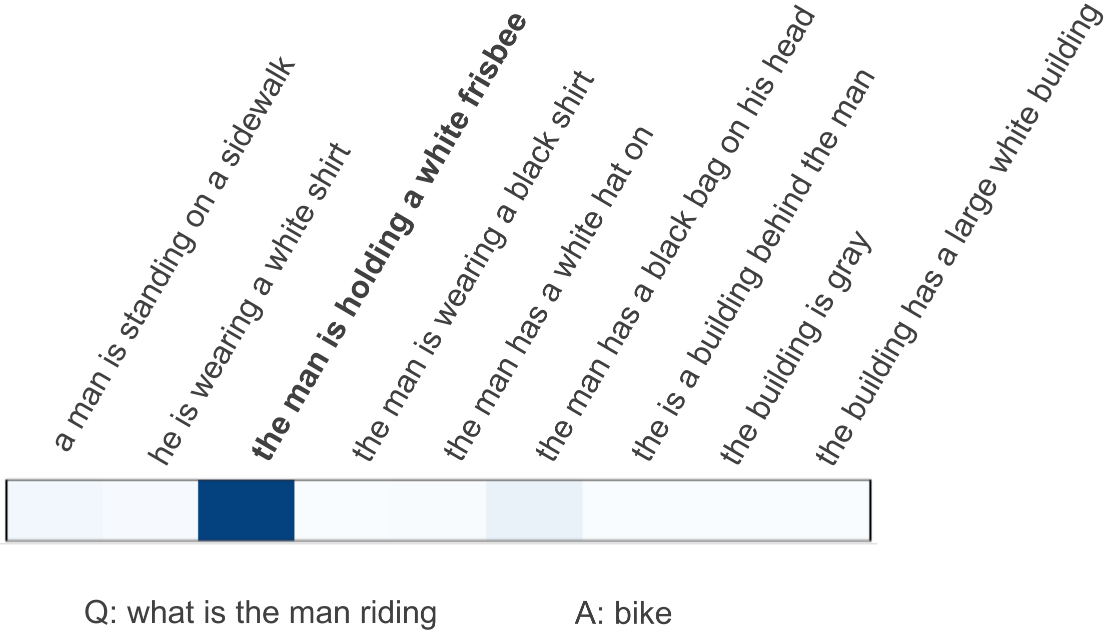


Q: what is the girl holding   A: bag

Q: what is the girl holding
Ground-truth A: bag
Model's A: suitcase

The paragraph misleads the model to the wrong answer
→ "the woman is holding a **suitcase**"

Attention Visualization (Failed Case)



Q: what is the man riding
Ground-truth A: bike
Model's A: frisbee

The paragraph misleads the model to the nearest wrong answer
→ "the man is holding a white **frisbee**"

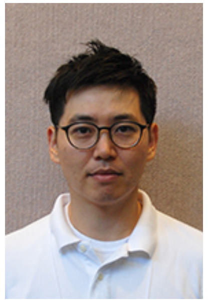
(slides by Hyounghun Kim)

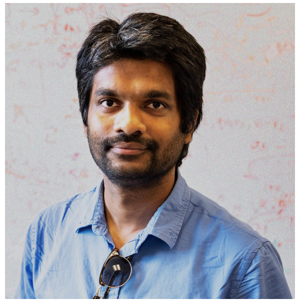
Video QA

(slides by Hyounghun Kim)

Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA

UNC NLP

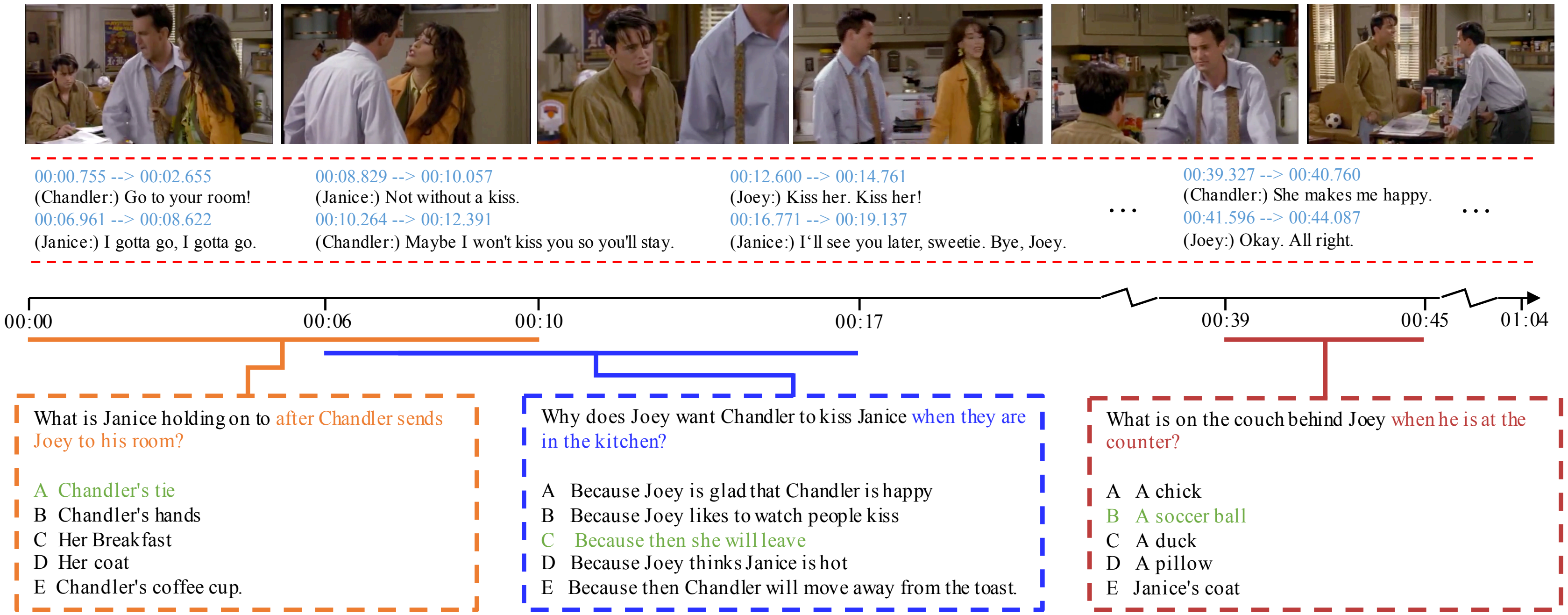




Hyounghun Kim   Zineng Tang   Mohit Bansal

ACL 2020

Task / Dataset



[1]

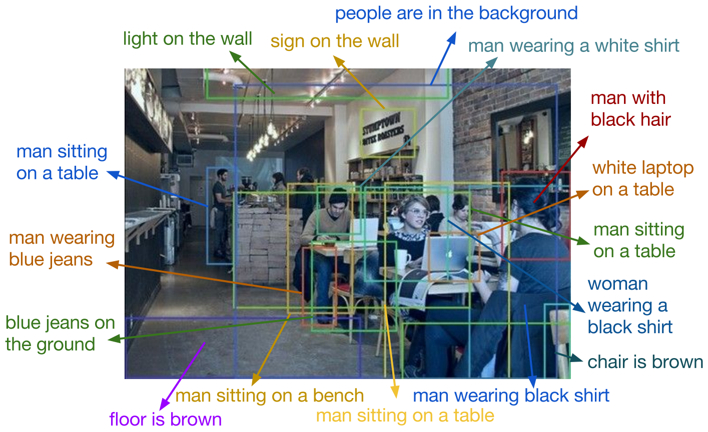
- We explore the TVQA dataset in this paper.
- The TVQA dataset consists of question and answer pairs, video frames, and corresponding subtitles.
- The task is to choose the correct answer among 5 candidates.

1. Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. TVQA: Localized, compositional video question answering. In EMNLP.

Motivation

- We present 3 contributions to improve video+dialogue QA:

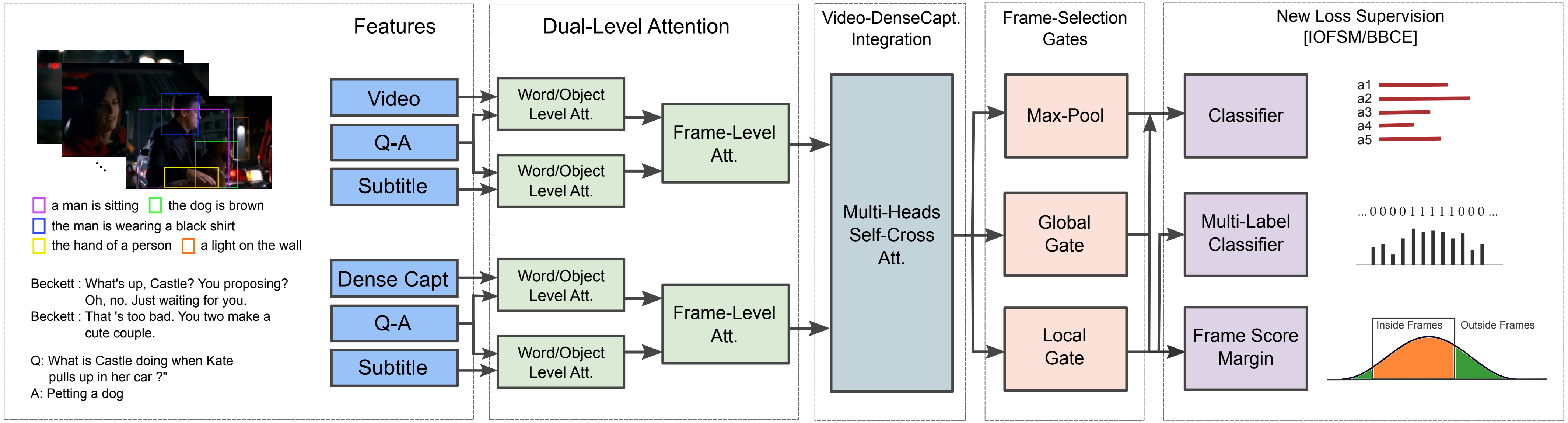
1. Dense-captions have diverse visual clues in the symbolic textual form. Thus, help provide cues for answering questions by matching/aligning keyword/phrase.



2. Frame gates pass relevant frames which has useful information for answering questions.

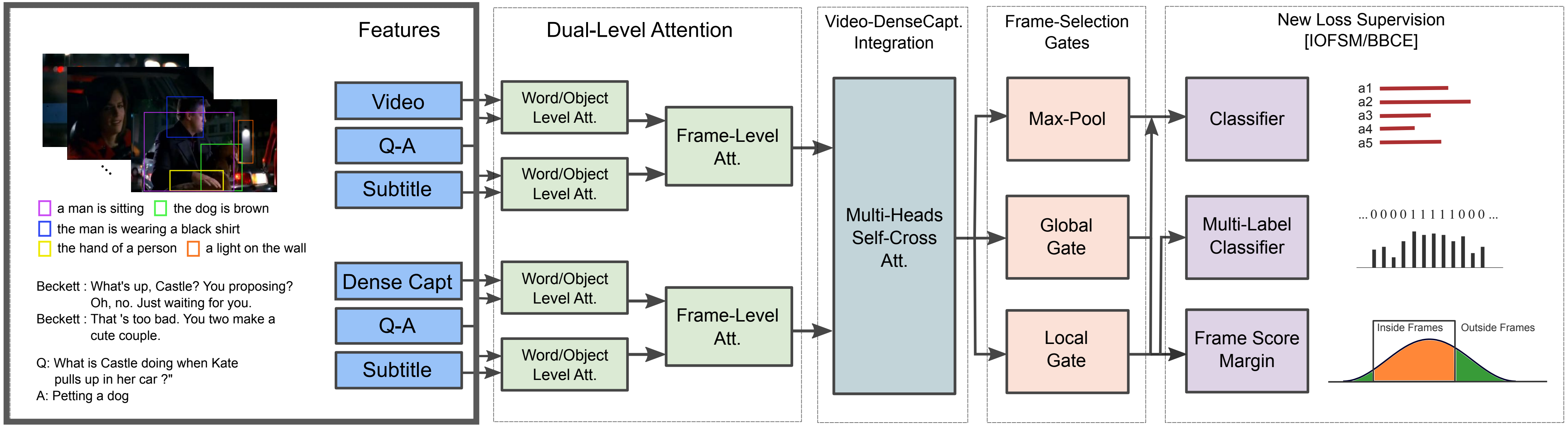
3. Localization/frame selection task can be cast as multi-label classification task and allow applying new customized losses.

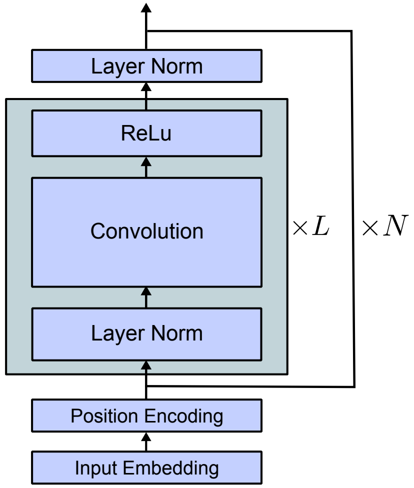
Model



Features
Video
Q-A
Subtitle
Dense Capt
Q-A
Subtitle
a man is sitting
the dog is brown
the man is wearing a black shirt
the hand of a person
a light on the wall
Beckett : What's up, Castle? You proposing?
Oh, no. Just waiting for you.
Beckett : That 's too bad. You two make a cute couple.
Q: What is Castle doing when Kate pulls up in her car ?"
A: Petting a dog
Dual-Level Attention
Word/Object Level Att.
Word/Object Level Att.
Frame-Level Att.
Word/Object Level Att.
Word/Object Level Att.
Frame-Level Att.
Video-DenseCapt. Integration
Multi-Heads Self-Cross Att.
Frame-Selection Gates
Max-Pool
Global Gate
Local Gate
New Loss Supervision [IOFSM/BBCE]
Classifier
a1
a2
a3
a4
a5
Multi-Label Classifier
...0000111111000...
Frame Score Margin
Inside Frames
Outside Frames

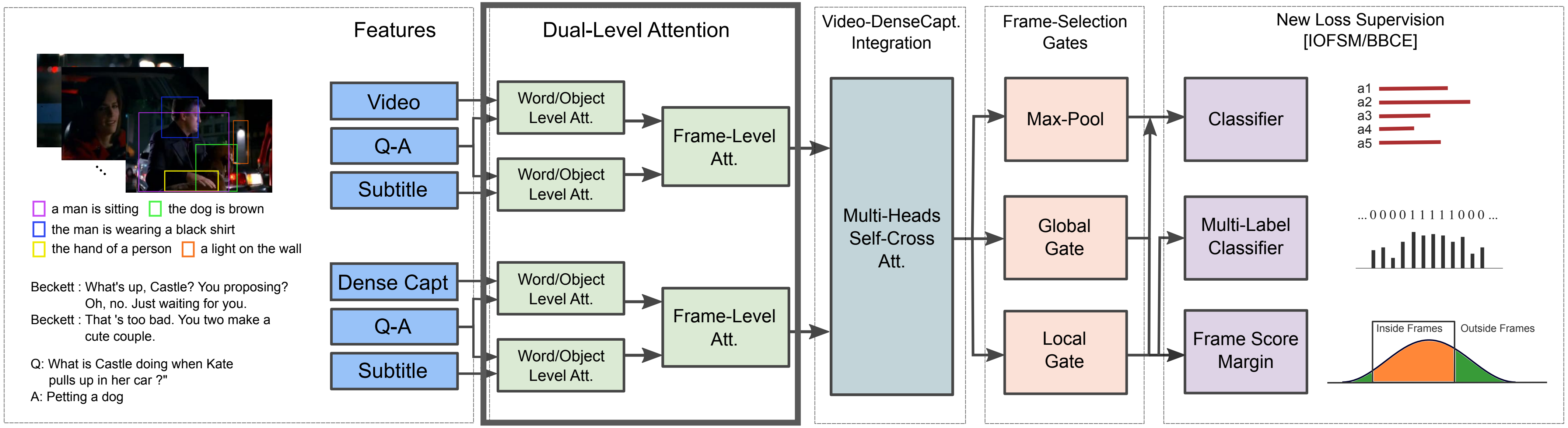
Model

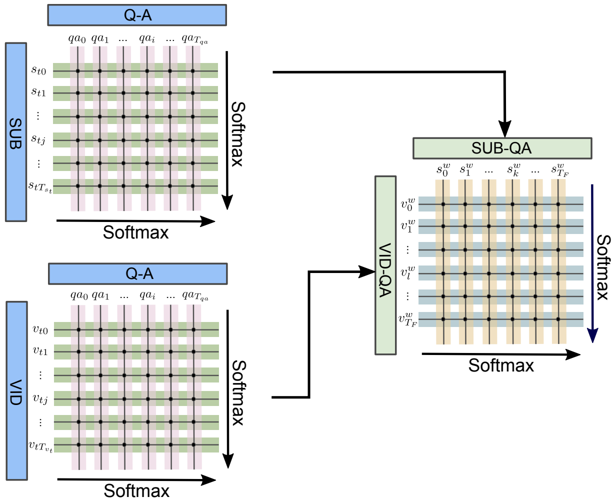




- The question and each of the 5 answer candidates are concatenated to create 5 QA pairs.
- Subtitles and dense captions are aligned with each frame which is extracted at 0.5 fps.
- All the input features are encoded with the convolutional layer.

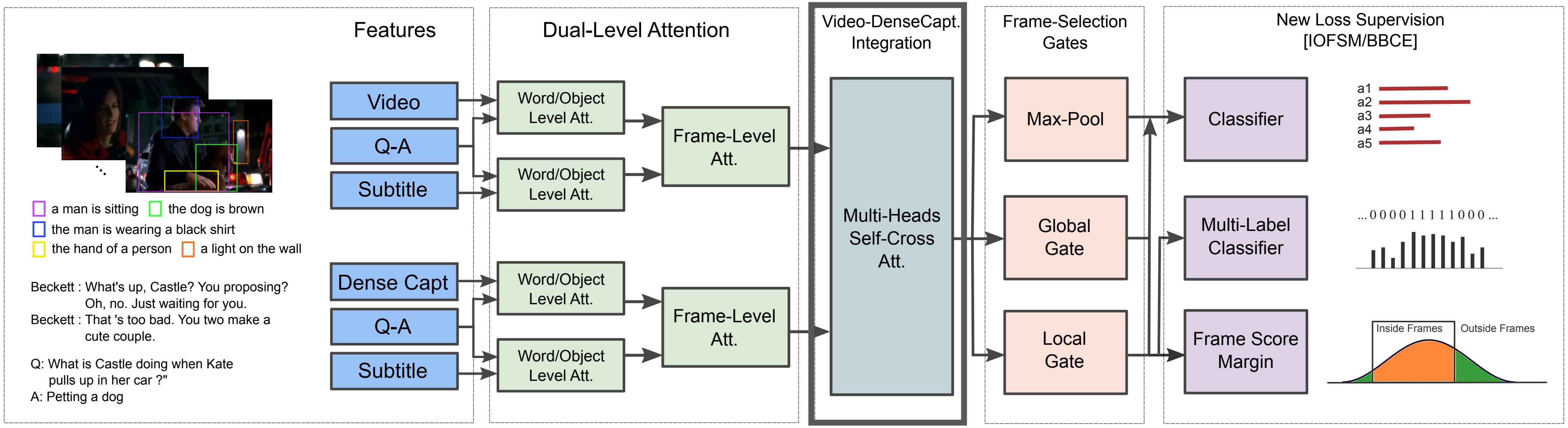
Model

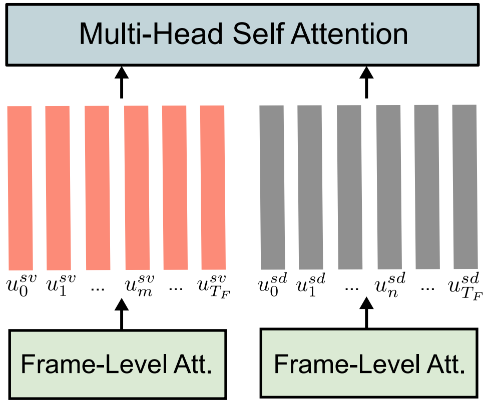




- First, the QAs are aligned with subtitle and video in the word/object level.
- Next, the fused features are aligned again in the frame level.
- The same dual-level attention is done with dense captions in the place of the video feature.

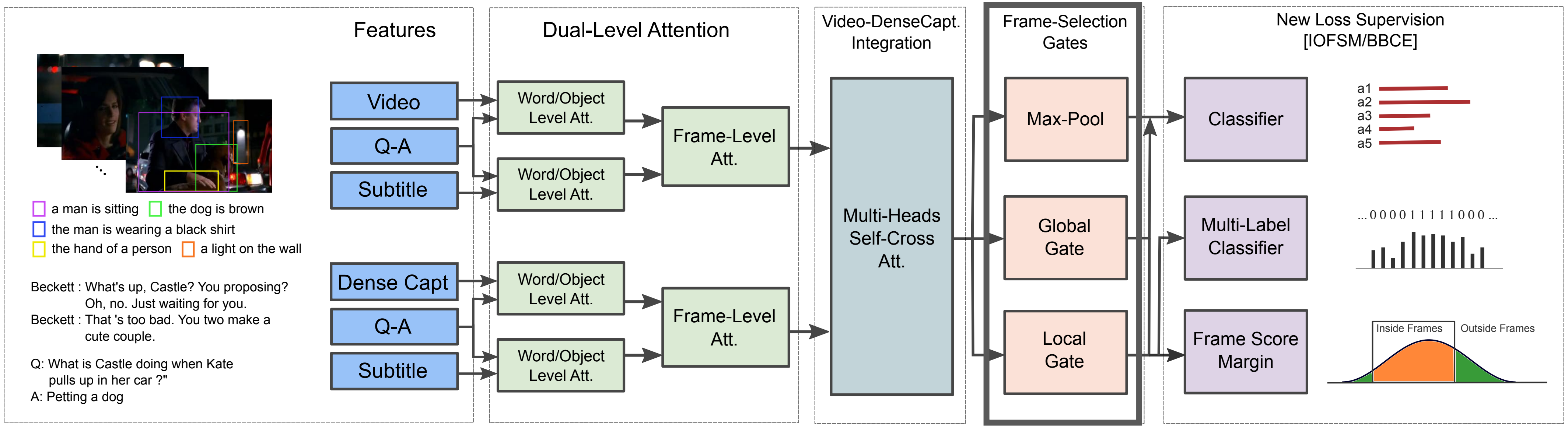
Model





- After the dual-level attention, we obtain two fused features: QA-SUB-VID and QA-SUB-DENSE.
- They are complementary, so need to be integrated.
- We use multi-head self-attention to combine them.

Model



$$\hat{z} = \phi_{\text{en2}}(z)$$

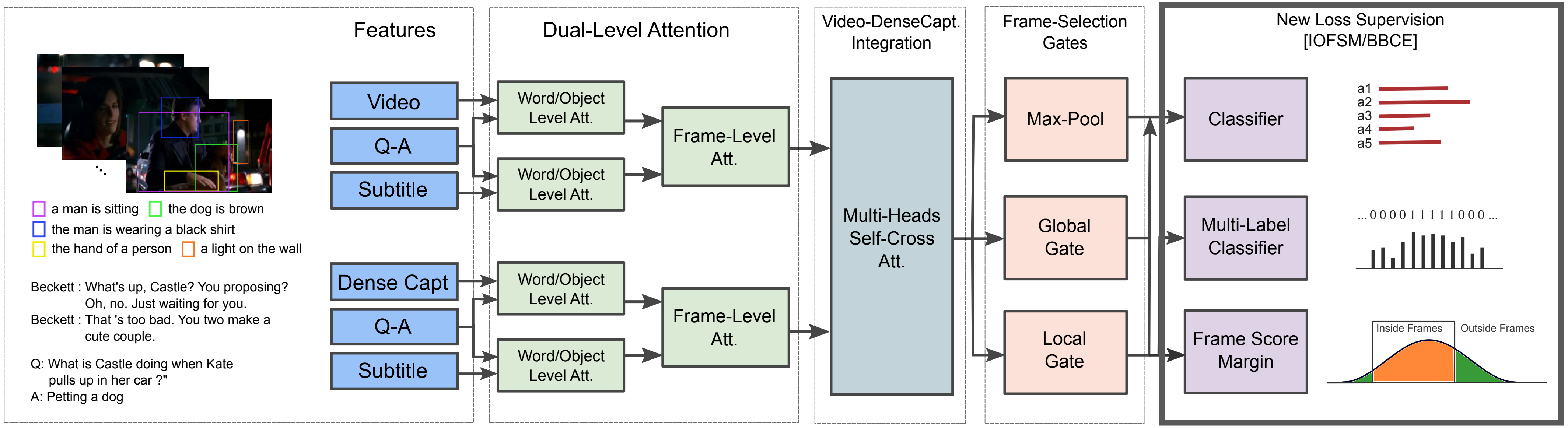
$$g^L = \text{sigmoid}(f^L(\hat{z}))$$

$$z^{gl} = \hat{z} \odot g^L$$

- The integrated feature still needs to be filtered to get more relevant information and select useful spatial-temporal information.
- Thus, frame-selection gates are applied.
- We use max-pooling, global gate, and local gate which is supervised by human importance annotations.

(slides by Hyounghun Kim)

Model



IOFSM: $loss_{io} = 1 + \mathrm{Avg}(\mathrm{OFS}) - \mathrm{Avg}(\mathrm{IFS})$

BBCE: $$loss_{bbce} = -\Big(\sum_{i}^{T_{F_{in}}} \log(s_i^{f_{in}})/T_{F_{in}} + \sum_{j}^{T_{F_{out}}} \log(1 - s_j^{f_{out}})/T_{F_{out}}\Big)$$

- To give higher weights to the relevant frames (in-frame) and lower weights to the less important frames (out-frame), we introduce new loss, In-and-out frame score margin (IOFSM).
- Also, for more balanced training signal, we introduce Balanced Binary Cross-Entropy (BBCE).

Results

UNC NLP

	Model	Test-Public (%)							Val (%)
		all	bbt	friends	himym	grey	house	castle	
1	jacobssy (anonymous)	66.01	68.75	64.98	65.08	69.22	66.45	63.74	64.90
2	multi-stream (Lei et al., 2018)	66.46	70.25	65.78	64.02	67.20	66.84	63.96	65.85
3	PAMN (Kim et al., 2019b)	66.77	-	-	-	-	-	-	66.38
4	Multi-task (Kim et al., 2019a)	67.05	-	-	-	-	-	-	66.22
5	ZGF (anonymous)	68.77	-	-	-	-	-	-	68.90
6	STAGE (Lei et al., 2020)	70.23	-	-	-	-	-	-	70.50
7	akalsdnr (anonymous)	70.52	71.49	67.43	72.22	70.42	70.83	72.30	71.13
8	Ours (hstar)	**74.09**	**74.04**	**73.03**	**74.34**	**73.44**	**74.68**	**74.86**	**74.20**

- Our model outperforms the state-of-the-art models by a large margin on both validation and test-public splits.

- Also, our model's scores across the TV shows are more balanced than any other models.

Results

	Model	Test-Public (%)							Val (%)
		all	bbt	friends	himym	grey	house	castle	
1	jacobssy (anonymous)	66.01	68.75	64.98	65.08	69.22	66.45	63.74	64.90
2	multi-stream (Lei et al., 2018)	66.46	70.25	65.78	64.02	67.20	66.84	63.96	65.85
3	PAMN (Kim et al., 2019b)	66.77	-	-	-	-	-	-	66.38
4	Multi-task (Kim et al., 2019a)	67.05	-	-	-	-	-	-	66.22
5	ZGF (anonymous)	68.77	-	-	-	-	-	-	68.90
6	STAGE (Lei et al., 2020)	70.23	-	-	-	-	-	-	70.50
7	akalsdnr (anonymous)	70.52	71.49	67.43	72.22	70.42	70.83	72.30	71.13
8	Ours (hstar)	**74.09**	**74.04**	**73.03**	**74.34**	**73.44**	**74.68**	**74.86**	**74.20**

- Our model outperforms the state-of-the-art models by a large margin on both validation and test-public splits.
- Also, our model's scores across the TV shows are more balanced than any other models.

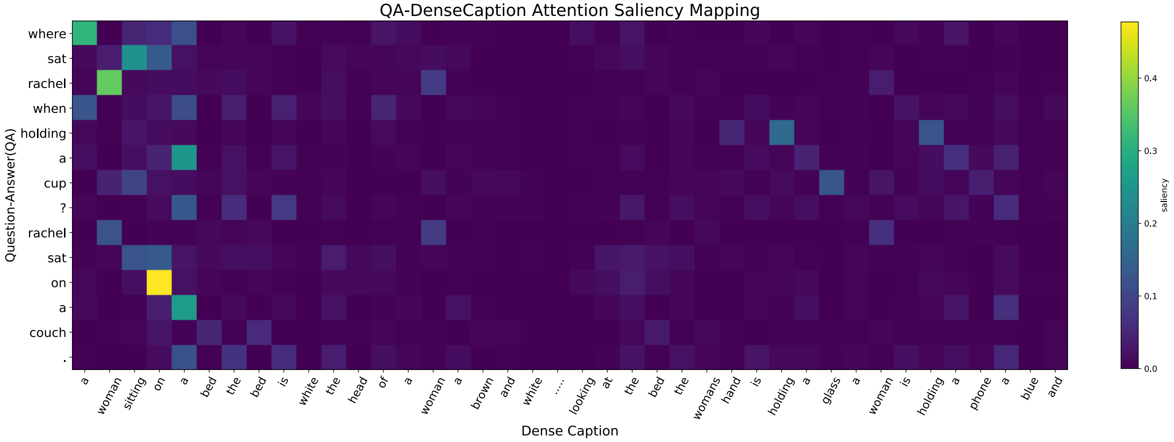
Ablation

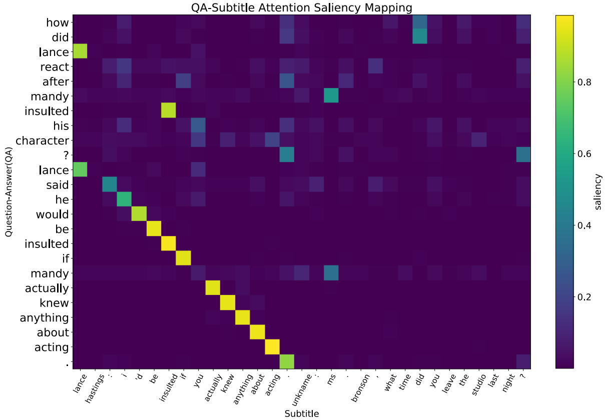
	Model	Val Score (%)
1	Single-Att + Frame-Span	69.86
2	Single-Att + Frame-Selection Gates	70.08
3	Dual-Att + Frame-Span	70.20
4	Dual-Att + Frame-Selection Gates (w/o NewLoss)	71.26
5	Dual-Att + Frame-Selection Gates	72.51
6	Dual-Att + Frame-Selection Gates (w/o NewLoss) + RoBERTa	72.53
7	Dual-Att + Frame-Selection Gates + RoBERTa	73.34
8	Dual-Att + Frame-Selection Gates + RoBERTa + DenseCapts	74.20

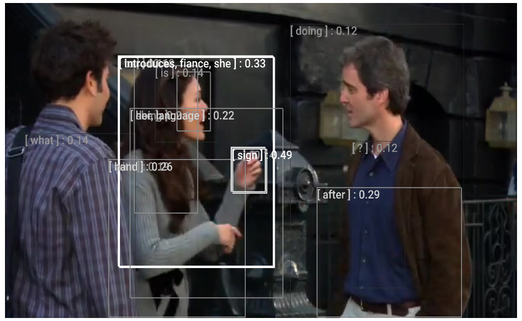
	Loss	Val Score (%)	IFS		OFS	
			avg	std	avg	std
1	BCE	71.26	0.468	0.108	0.103	0.120
2	IOFSM	70.75	0.739	0.127	0.143	0.298
3	BCE+IOFSM	72.22	0.593	0.128	0.111	0.159
4	BBCE	72.27	0.759	0.089	0.230	0.231
5	BBCE+IOFSM	72.51	0.764	0.098	0.182	0.246

- Each component of our model helps increase performance.
- Especially, our new losses, OFSM and BBCE (row 5 vs 4: $p < 0.0001$, row 7 vs 6: $p < 0.005$), and dense captions (row 8 vs 7: $p < 0.005$) improve performance significantly.
- Using IOFSM alone decreases the score by increasing OFS std.
- Using IOFSM+BBCE increases the score by increasing avg. IFS.

Visualization (word/object level att.)



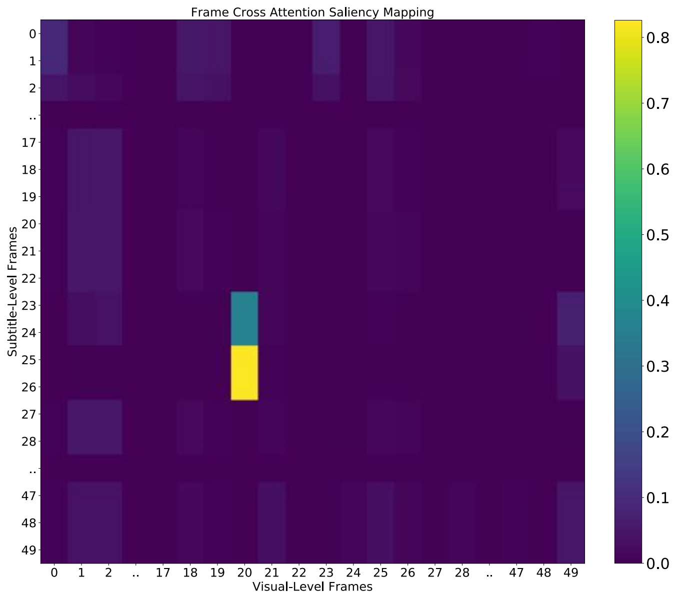




Q: What is Cathy doing with her hand after she introduces her fiance to Ted?
A: She is doing sign language.

- Dense captions help localize the relevant frame by matching keyword/phrase (e.g., “a woman sitting”, “holding a glass”).
- Subtitles also help answer the question by providing a nearly exact clue for the answer (i.e., “... anything about acting.”).
- Object level attention helps by aligning the word in the QA and the object feature in a video frame (i.e., the woman’s hand and ‘sign’ in the QA).

Visualization (frame level att.)



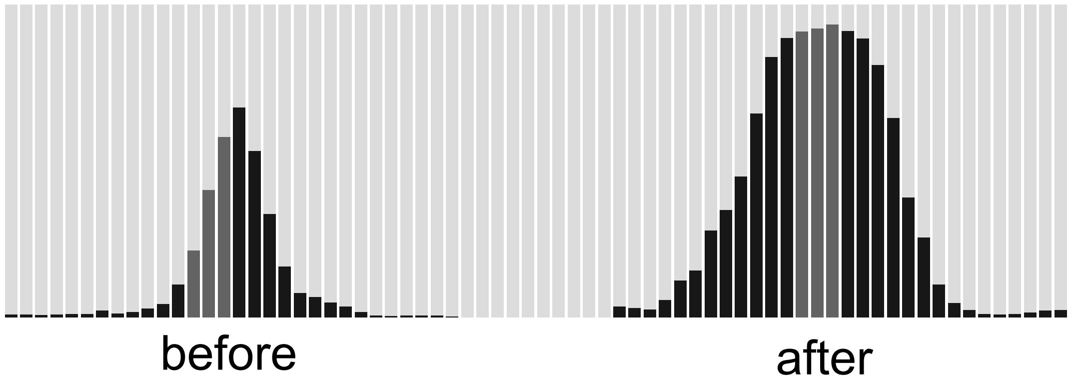
Esposito : Upstairs. go.
Unkname : Carol!

Frame 20 Frame 25

Q: Where did Esposito search after he searched Carol 's house downstairs?
A: Upstairs.

- Frame-level attention can align relevant frames from different features.
- In the example, to answer the question, the model needs to find a frame in which ‘he (Esposito) searched Carol’s house downstairs’, then find a frame which has a clue for ‘where did Esposito search’. Our frame-level attention can properly align the information fragments from different features (Frame 20 and 25) to help answer questions.

Visualization (new losses)



- Our new losses (IOFSM+BBCE) changes the score distribution over all frames.
- Before applying our losses (left side), overall scores are relatively low. After using the losses (right side), overall scores are increased, and especially, scores around in-frames get much higher.

(slides by Hyounghun Kim)

Visual Dialog

(slides by Hyounghun Kim)
AAAI 2020

Modality-Balanced Models for Visual Dialogue

Hyounghun Kim, Hao Tan, Mohit Bansal

University of North Carolina at Chapel Hill

Model Bias for VisDial: We show that models have different behaviors on VisDial (Das et al. 2017) when being evaluated on different metrics (NDCG, MMR, recall@k, etc.).

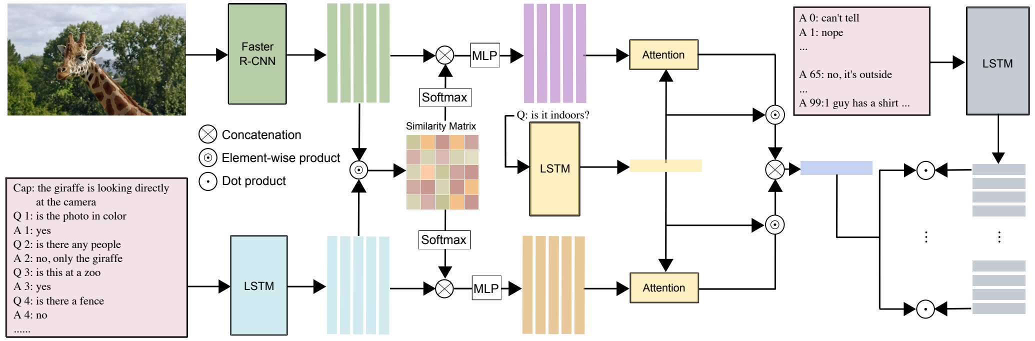
Image-Only Model: Our image-only model gives higher NDCG score which could imply that image-only model is better at generalization.

Image-History Joint Model: Our image-history joint model has higher non-NDCG (MMR, recall@k, etc.) which could imply that image-history model is good at keyword matching / memorizing to give accurate answers.

Combined Fusion Model: Explicitly maintain two models; complementary abilities for a more balanced multimodal model.

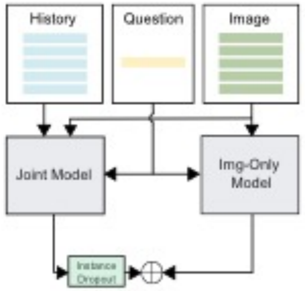
Models	NDCG	MRR	R@1	R@5	R@10	Mean
FULL	57.81	**64.47**	**50.87**	**81.38**	90.03	**4.10**
H-5	58.24	64.29	50.61	81.35	**90.22**	4.10
H-1	59.29	62.86	49.07	79.76	89.08	4.35
Img-only	**61.04**	61.25	47.18	78.43	88.17	4.61

- Performance of models with different amounts of history.
- The more history the less NDCG (vice versa).



Consensus Dropout Fusion (CDF):

- Dropout the final features from the joint model randomly.
- Modulates the influence of the joint model.



Models	NDCG	MRR	R@1	R@5	R@10	Mean
Img-Only	61.04	61.25	47.18	78.43	88.17	4.61
Joint	58.97	64.57	50.87	81.58	90.30	4.05
CDF	59.93	64.52	50.92	81.31	90.00	4.10
Ensemble	61.20	64.67	51.00	81.60	90.37	4.03

- CDF model shows more balanced results while the ensemble model takes the best from both models.

(slides by Hyounghun Kim)
AAAI 2020

Modality-Balanced Models for Visual Dialogue

Hyounghun Kim, Hao Tan, Mohit Bansal

University of North Carolina at Chapel Hill

	Models	NDCG	MRR	R@1	R@5	R@10	Mean
	LF (Das et al. 2017)	45.31	55.42	40.95	72.45	82.83	5.95
	HRE (Das et al. 2017)	45.46	54.16	39.93	70.45	81.50	6.41
	MN (Das et al. 2017)	47.50	55.49	40.98	72.30	83.30	5.92
	MN-att (Das et al. 2017)	49.58	56.90	42.43	74.00	84.35	5.59
	LF-att (Das et al. 2017)	49.76	57.07	42.08	74.83	85.05	5.41
	CorefNMN (Kottur et al. 2018)	54.7	61.5	47.55	78.10	88.80	4.40
Visual Dialog challenge 2018	RvA (Niu et al. 2018)	55.59	63.03	49.03	80.40	89.83	4.18
	USTC-YTH (Yang, Zha, and Zhang 2019)	57.17	64.22	50.88	80.63	89.45	4.20
	DL-61 (single) (Guo, Xu, and Tao 2019)	57.32	62.20	47.90	80.43	89.95	4.17
	DL-61 (ensemble) (Guo, Xu, and Tao 2019)	57.88	63.42	49.30	80.77	90.68	3.97
Visual Dialog challenge 2019	DAN (single) (Kang, Lim, and Zhang 2019)	57.59	63.20	49.63	79.75	89.35	4.30
	DAN (ensemble) (Kang, Lim, and Zhang 2019)	59.36	64.92	51.28	81.60	90.88	3.92
	ReDAN+ (ensemble) (Gan et al. 2019)	64.47	53.73	42.45	64.68	75.68	6.63
	MReaL–BDAI (not published)	74.02	52.62	40.03	65.85	79.15	6.76
	Our Image-Only (ensemble)	60.16	61.26	47.15	78.73	88.48	4.46
	Our Consensus Dropout Fusion (ensemble)	59.49	64.40	50.90	81.18	90.40	3.99

- Evaluation on test-standard dataset.

Models	NDCG	MRR	R@1	R@5	R@10	Mean
CDF (p=0.00)	59.40	64.61	51.01	81.73	90.30	4.06
CDF (p=0.15)	59.49	64.64	50.94	81.63	90.07	4.07
CDF (p=0.25)	59.93	64.52	50.92	81.31	90.00	4.10
CDF (p=0.35)	60.11	64.21	50.56	81.20	89.84	4.15

- Consensus dropout fusion and different dropout rates. With different dropout rates, consensus dropout fusion model yields different scores of all metrics.

Models	NDCG	MRR	R@1	R@5	R@10	Mean
Img+Img	61.97	62.24	48.20	79.49	88.83	4.41
Joint+Joint	59.84	65.60	52.06	82.46	90.87	3.88
Img+Joint	61.50	65.04	51.38	81.93	90.45	3.96

- Performance of ensemble models with different combinations. Img+Img model has the highest value of NDCG while Joint+Joint model has the highest values for other metrics. Img+Joint model has more balanced results.

Cap: the blue utility truck is parked in front of the gas station

Q1: what kind of truck is it
A1: concrete truck
Q2: is it a dark blue
A2: kind of like a sky blue
...
Q6: is there a store with the gas station
A6: yes
Q7: is it a big store
A7: no more like a **convenience store**

Q: what color is **it**
A: the building has a red awning and stone front

Cap: female tennis players stand on a tennis court

Q1: do you see any other people
A1: yes, other **3 persons**
Q2: what color is her hair
A2: not visible, covered by a hat
...
Q7: do they all have tennis racquets
A7: yes
Q8: is it day or night
A8: it is day

Q: do they have spectators
A: no, only **3 persons** playing and 1 watching

Q: can you see the color of the phone
A: yes, white

Q: what is the shape of the pizza
A: 2 triangle slices

(slides by Hyounghun Kim)

Thank you

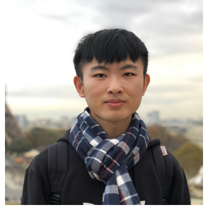
Outline

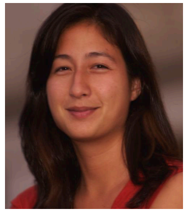
- Video Question Answering
  - TVQA: Localized, Compositional Video Question Answering, EMNLP 2018

- Language-driven Moment Localization
  - TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval, ECCV 2020

- Future Event Prediction
  - What is More Likely to Happen Next? Video-and-Language Future Event Prediction, EMNLP 2020

UNC NLP






Jie Lei, Licheng Yu, Tamara L. Berg, Mohit Bansal
UNC Chapel Hill

(Slides by Jie Lei)

TVQA: Localized, Compositional Video Question Answering

tvqa.cs.unc.edu

Jie Lei, Licheng Yu, Mohit Bansal, Tamara L. Berg

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

EMNLP 2018, Brussels, Belgium

(Slides by Jie Lei)

What is Sheldon holding when he is talking to Howard about the sword?

0) A comic book
1) A computer
2) A sword
3) A toy train
4) A drink

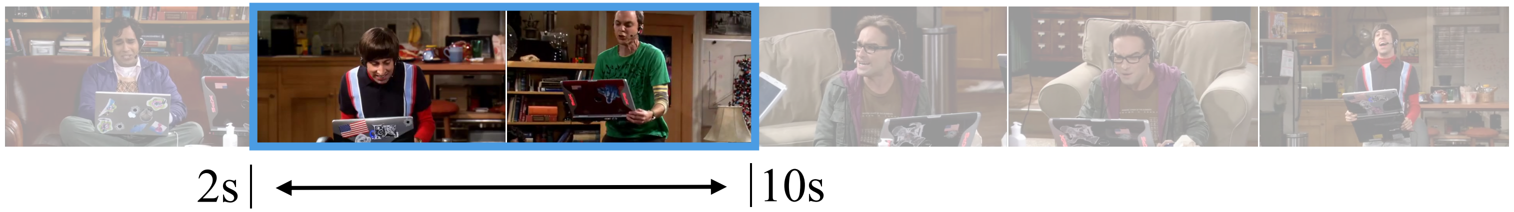
2s | ⟷ | 10s

(Slides by Jie Lei)

TVQA Dataset - Example

What is Sheldon holding when he is talking to Howard about the sword?

0) A comic book
1) A computer
2) A sword
3) A toy train
4) A drink



(Slides by Jie Lei)

Write a question:

[What/Why/...] ___ [when/before/after] ___

Question + Localization

What is Sheldon holding when he is talking to Howard about sword?

What is Sheldon holding when
he is talking to Howard about sword?

A computer

1) A comic book
2) A sword
3) A toy train
4) A drink

Mark the START and END timestamps:

2s | ← → | 10s

START END

(Slides by Jie Lei)

6 TV shows, 3 genres:

- **Sitcom**: *The Big Bang Theory, Friends, How I Met Your Mother*
- **Medical**: *Grey's anatomy, House M.D.*
- **Crime**: *Castle*

Show	Genre	#Sea.	#Epi.	#Clip	#QA
BBT	sitcom	10	220	4,198	29,384
Friends	sitcom	10	226	5,337	37,357
HIMYM	sitcom	5	72	1,512	10,584
Grey	medical	3	58	1,427	9,989
House	medical	8	176	4,621	32,345
Castle	crime	8	173	4,698	32,886
Total	—	44	925	21,793	152,545

Data Statistics for each TV show. BBT = *The Big Bang Theory*, HIMYM = *How I Met You Mother*, Grey = *Grey's Anatomy*, House = *House M.D.*, Epi = Episode, Sea. = Season

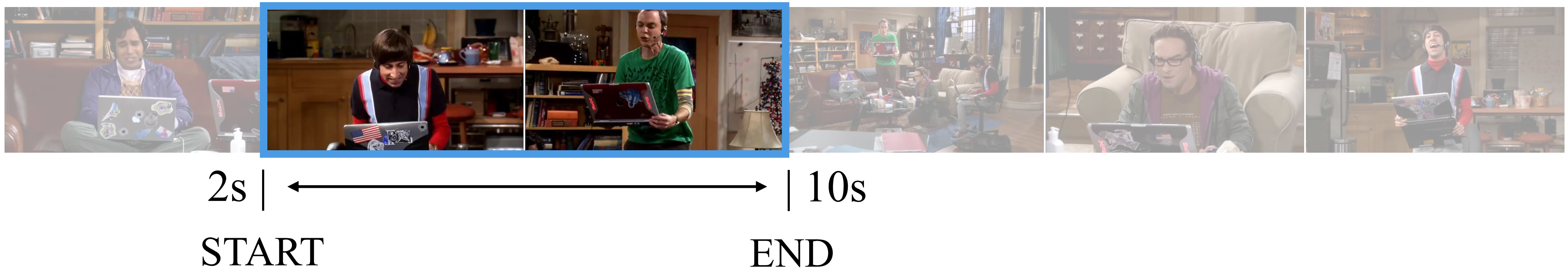
Different show comes with different words

Show	Top unique nouns
BBT	game, mom, laptop, water, store, dinner, book, stair, computer, food, wine, glass, couch, date
Friends	shop, kiss, hair, sofa, jacket, counter, coffee, everyone, coat, chair, kitchen, baby, apartment
HIMYM	bar, beer, drink, job, dad, sex, restaurant, wedding, party, booth, dog, story, bottle, club, painting
Grey	nurse, side, father, hallway, scrub, chart, wife, window, life, family, chief, locker, head, surgery
House	cane, team, blood, test, brain, pill, office, pain, symptom, diagnosis, hospital, coffee, cancer, drug
Castle	gun, victim, picture, case, photo, body, murder, suspect, scene, crime, money, interrogation

Top unique nouns in question and correct answer

- Task 1: Question answering without timestamp annotation

- Task 2: Question answering with timestamp annotation



Model - Input

What is Sheldon holding when he is talking to Howard about sword?

0) A comic book
1) A computer
2) A sword
3) A toy train
4) A drink

A Question with
5 candidate answers

(Howard:)Sheldon, he's got Raj. Use your sleep spell. Sheldon. Sheldon.

(Sheldon:)I've got the Sword of Azeroth.
…

Subtitle sentences

Video frames

(Slides by Jie Lei)

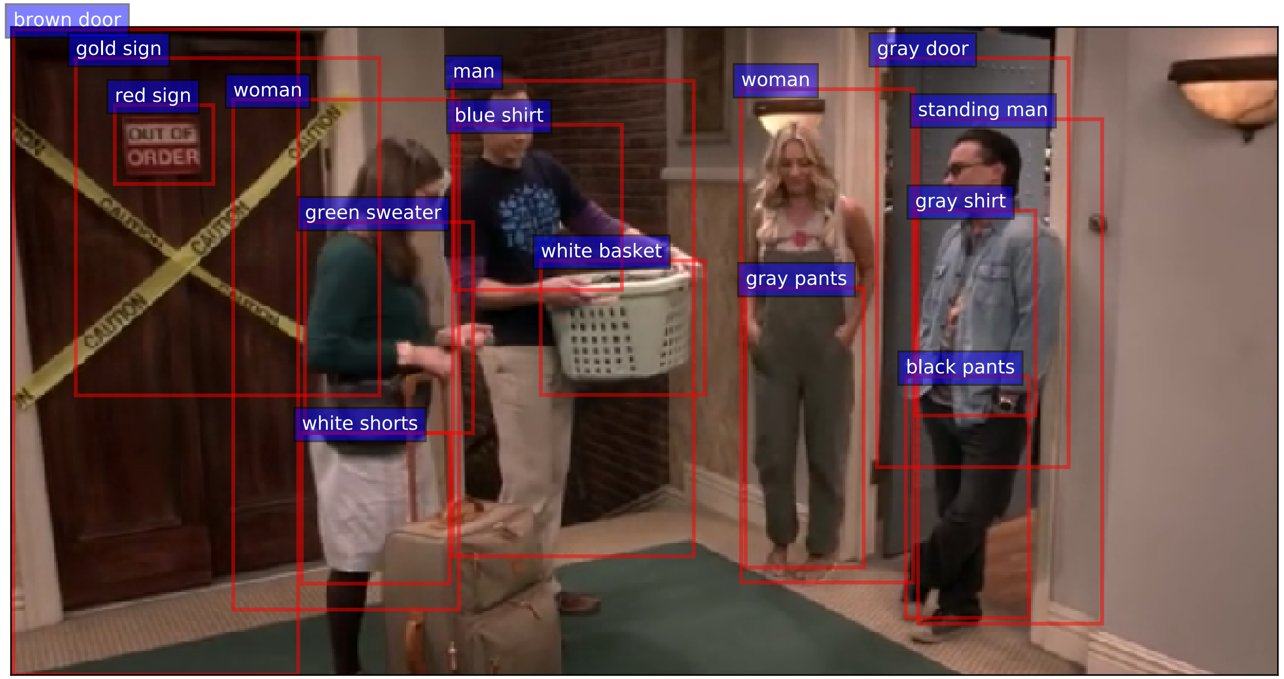
→ ResNet101 → ImageNet feature (img)

→ Object Detector → Regional visual feature (reg)

Visual concepts feature (cpt)

Faster R-CNN trained on Visual Genome

He, Kaiming, et al. Deep residual learning for image recognition. CVPR 2016
Russakovsky, Olga, et al. Imagenet large scale visual recognition challenge. IJCV 2015



brown door, gold sign, red sign, woman, white shorts, green sweater, man, blue shirt, white basket, woman, gray pants, gray door, standing man, gray shirt, black pants

Faster R-CNN detection example

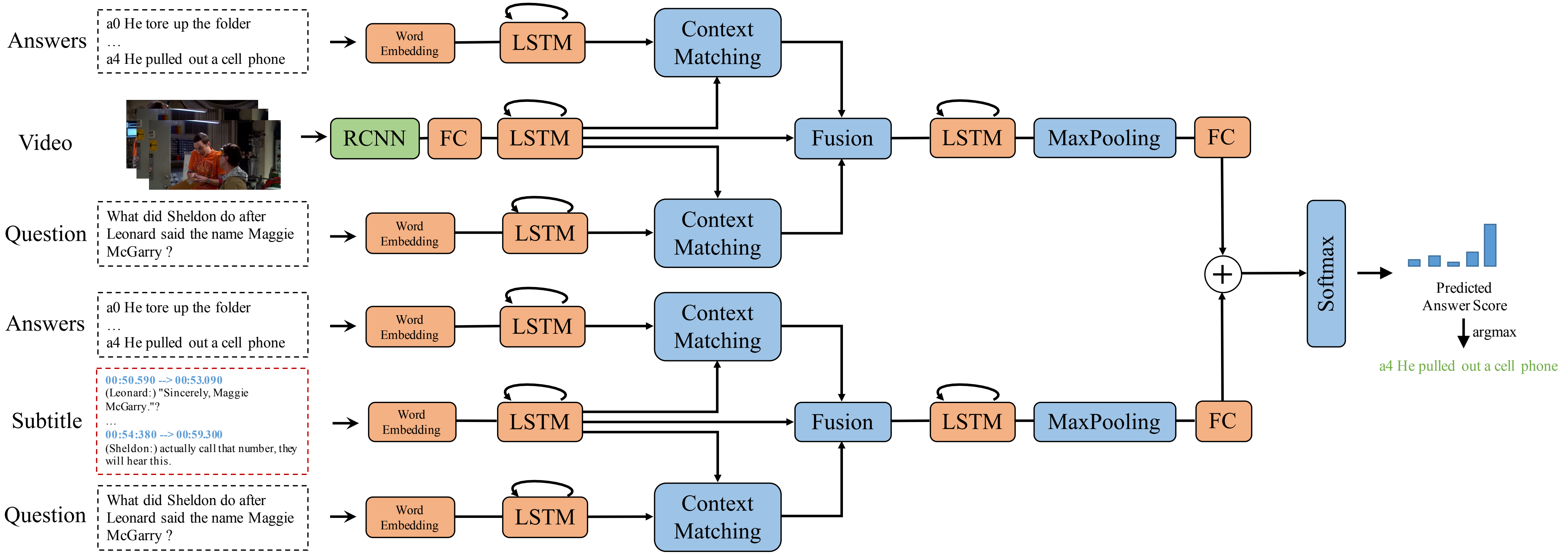
The regional visual feature (image embeddings inside the bounding boxes) and visual concepts feature (shown in the caption) can be used to answer the question:

'What is Sheldon holding when everyone is at the door?' (basket).

(Slides by Jie Lei)

Ren, Shaoqing, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015
Anderson, Peter, et al. Bottom-up and top-down attention for image captioning and VQA. *CVPR 2018*
Krishna, Ranjay, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV 2017*

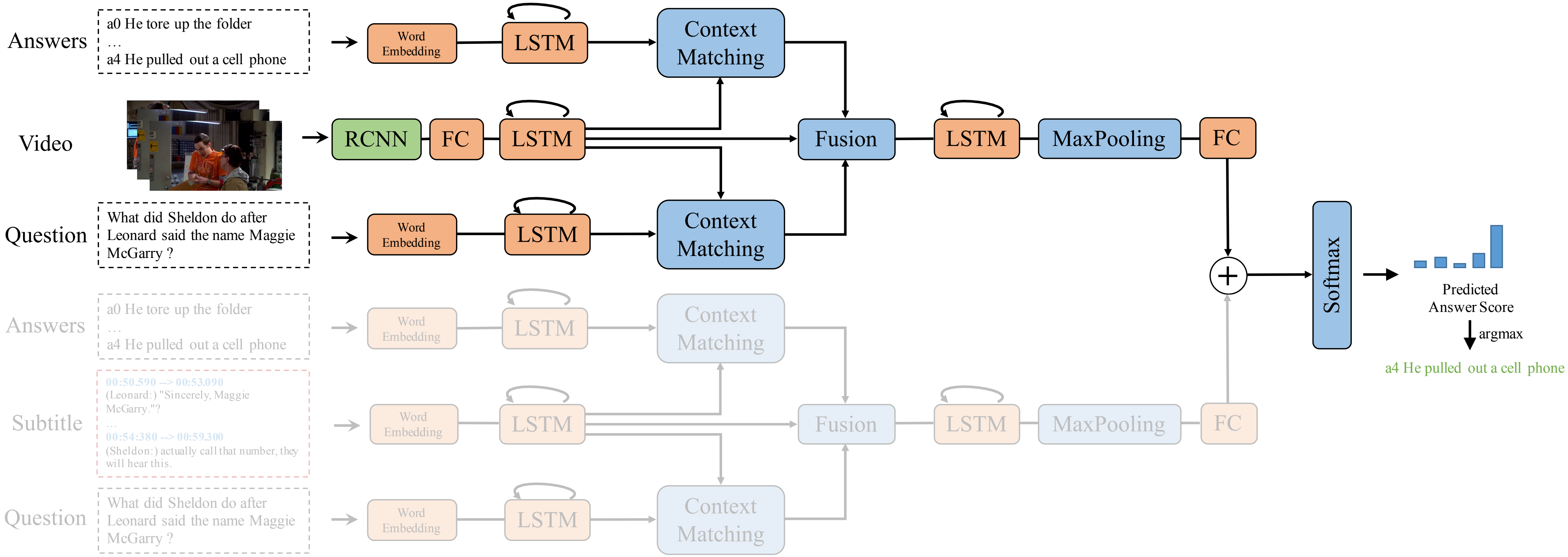
Multiple streams, each stream deals with different contextual input



(Slides by Jie Lei)

Multiple streams, each stream deals with different contextual input

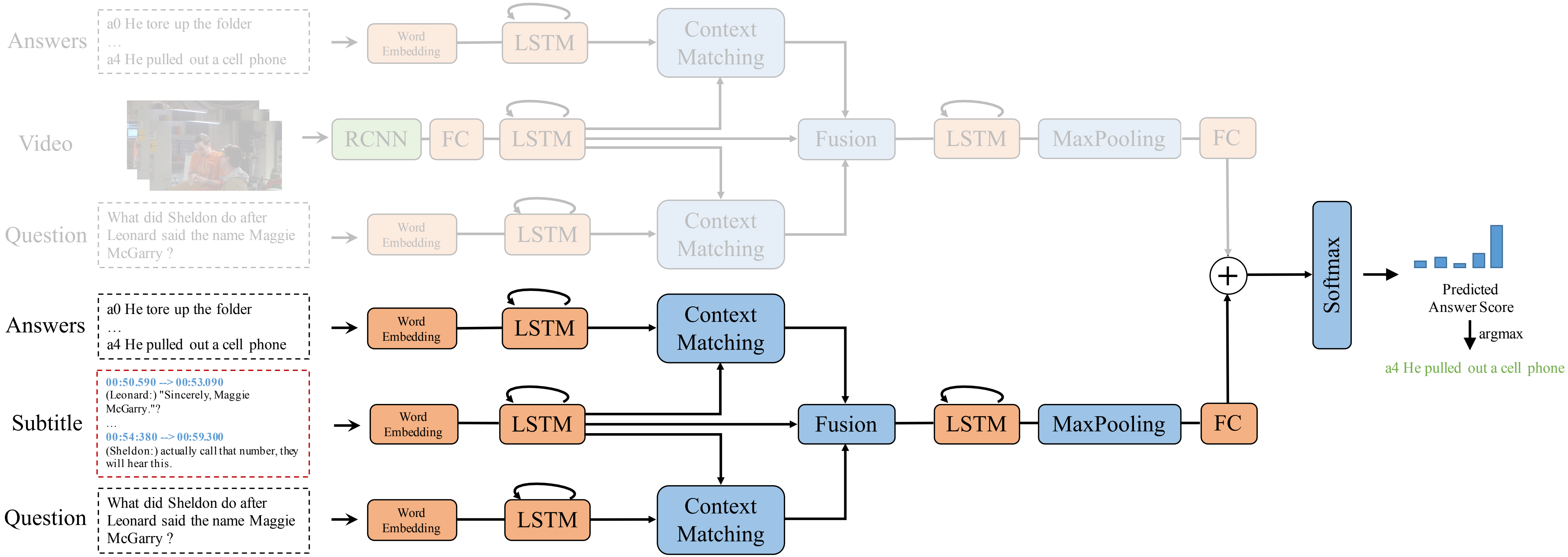
- Upper stream -- Video



(Slides by Jie Lei)

Multiple streams, each stream deals with different contextual input

- Bottom stream -- Subtitle



	Method	Video Feature	Test Accuracy w/o ts	Test Accuracy w/ ts
0	Random	-	20.00	20.00
1	Longest Answer	-	30.41	30.41
2	Retrieval-Glove	-	22.48	22.48
3	Retrieval-SkipThought	-	24.24	24.24
4	Retrieval-TFIDF	-	20.88	20.88
5	NNS-Glove Q	-	22.40	22.40
6	NNS-SkipThought Q	-	23.79	23.79
7	NNS-TFIDF Q	-	20.33	20.33
8	NNS-Glove S	-	23.73	29.66
9	NNS-SkipThought S	-	26.81	37.87
10	NNS-TFIDF S	-	49.94	51.23
11	Our Q	-	43.34	43.34
12	Our V+Q	img	42.67	43.69
13	Our V+Q	reg	42.75	44.85
14	Our V+Q	cpt	43.38	45.41
15	Our S+Q	-	63.14	66.23
16	Our S+V+Q	img	63.57	66.97
17	Our S+V+Q	reg	63.19	67.82
18	Our S+V+Q	cpt	**65.46**	**68.60**

Question only

Add Video

Add Subtitle

Add Video, Subtitle

Accuracy for different methods on TVQA test set. Q = Question, S = Subtitle, V = Video, img = ImageNet features, reg = regional visual features, cpt = visual concept features, ts = timestamp annotation.

Both visual and textual information are important!

(Slides by Jie Lei)

	Method	Video Feature	Test Accuracy w/o ts	Test Accuracy w/ ts
0	Random	-	20.00	20.00
1	Longest Answer	-	30.41	30.41
2	Retrieval-Glove	-	22.48	22.48
3	Retrieval-SkipThought	-	24.24	24.24
4	Retrieval-TFIDF	-	20.88	20.88
5	NNS-Glove Q	-	22.40	22.40
6	NNS-SkipThought Q	-	23.79	23.79
7	NNS-TFIDF Q	-	20.33	20.33
8	NNS-Glove S	-	23.73	29.66
9	NNS-SkipThought S	-	26.81	37.87
10	NNS-TFIDF S	-	49.94	51.23
11	Our Q	-	43.34	43.34
12	Our V+Q	img	42.67	43.69
13	Our V+Q	reg	42.75	44.85
14	Our V+Q	cpt	43.38	45.41
15	Our S+Q	-	63.14	66.23
16	Our S+V+Q	img	63.57	66.97
17	Our S+V+Q	reg	63.19	67.82
18	Our S+V+Q	cpt	**65.46**	**68.60**

Accuracy for different methods on TVQA test set. Q = Question, S = Subtitle, V = Video, img = ImageNet features, reg = regional visual features, cpt = visual concept features, ts = timestamp annotation.

Timestamp information is helpful

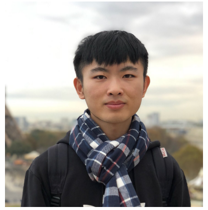
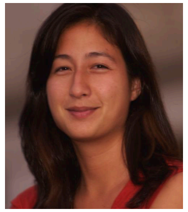
(Slides by Jie Lei)

TVQA dataset and code:

https://tvqa.cs.unc.edu/

https://github.com/jayleicn/TVQA

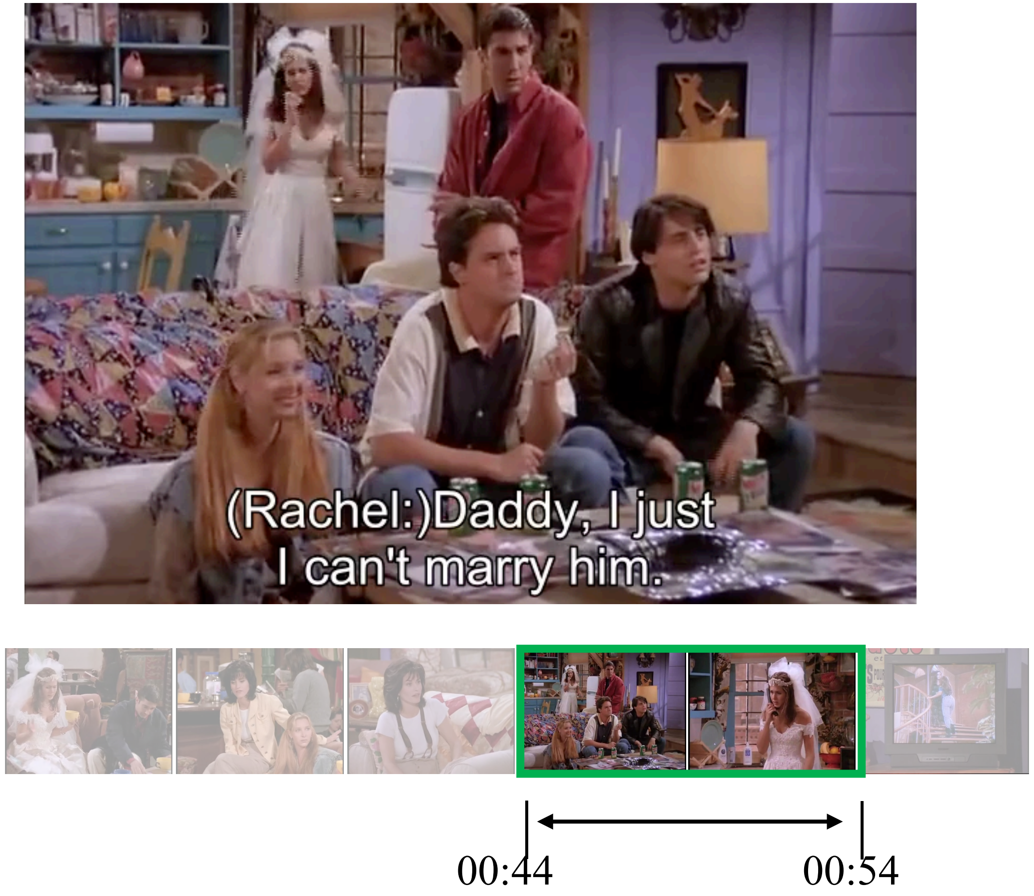
TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval

Jie Lei, Licheng Yu, Tamara L. Berg, Mohit Bansal
UNC Chapel Hill

(Slides by Jie Lei)

TVR Moment Retrieval Example



Query: Rachel explains to her dad on the phone why she can't marry her fiancé.
Query Type: video + subtitle

(Slides by Jie Lei)

TVR Moment Retrieval Example

00:44 00:54

Query: Rachel explains to her dad on the phone why she can't marry her fiancé.

Query Type: video + subtitle

(Slides by Jie Lei)

TVR Data Ananlysis

Percentage of queries that have multiple actions or involve multiple people. We also show query examples, with unique person mentions <u>underlined</u> and actions in **bold**

Dataset	#actions ≥2 (%)	#people ≥2 (%)	Query examples (*query type*)
TACoS [28]	20	0	<u>She</u> **rinses** the peeled carrots off in the sink. (*video*) The <u>person</u> **removes** roots and outer leaves and **rewashes** the leek. (*video*)
CharadesSTA [8]	6	12	A <u>person</u> is **eating** food slowly. (*video*) A <u>person</u> is **opening** the door to a bedroom. (*video*)
ActivityNet Caption [21]	44	44	<u>He</u> then **grabs** a metal mask and **positions** himself correctly on the floor. (*video*) The same <u>man</u> **comes** back and **lifts** the weight over his head again. (*video*)
DiDeMo [13]	6	10	A dog **shakes** its body. (*video*) A <u>lady</u> in a cowboy hat **claps** and **jumps** excitedly. (*video*)
TVR	67	66	<u>Bert</u> **leans** down and **gives** <u>Amy</u> a hug who is **standing** next to <u>Penny</u>. (*video*) <u>Taub</u> **argues** with the <u>patient</u> that fighting in Hockey **undermines** the sport. (*sub*) <u>Chandler</u> **points** at <u>Joey</u> while **describing** a <u>woman</u> who wants to **date** him. (*video+sub*)

- Compared to existing datasets, TVR queries typically have more people and actions and require both video and sub (subtitle) context.

(Slides by Jie Lei)

TVR Task

Video Corpus Moment Retrieval (VCMR)

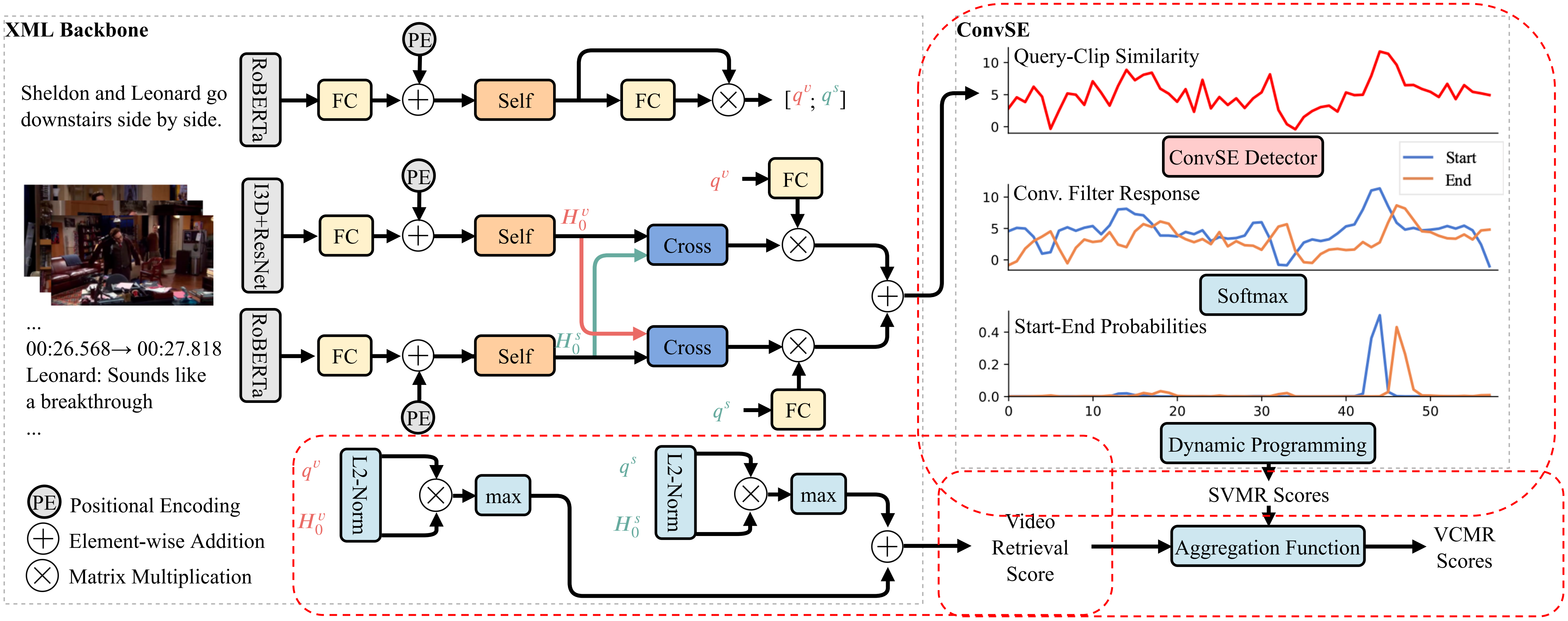
- A query + A video corpus ➡ Retrieve the matched moment from the corpus.
  - Retrieve the GT video. (Video Retrieval)
  - Localize the moment from the retrieved video. (Single Video Moment Retrieval)

Query: Rachel explains to her dad on the phone why she can't marry her fiancé.

Query Type: video + subtitle

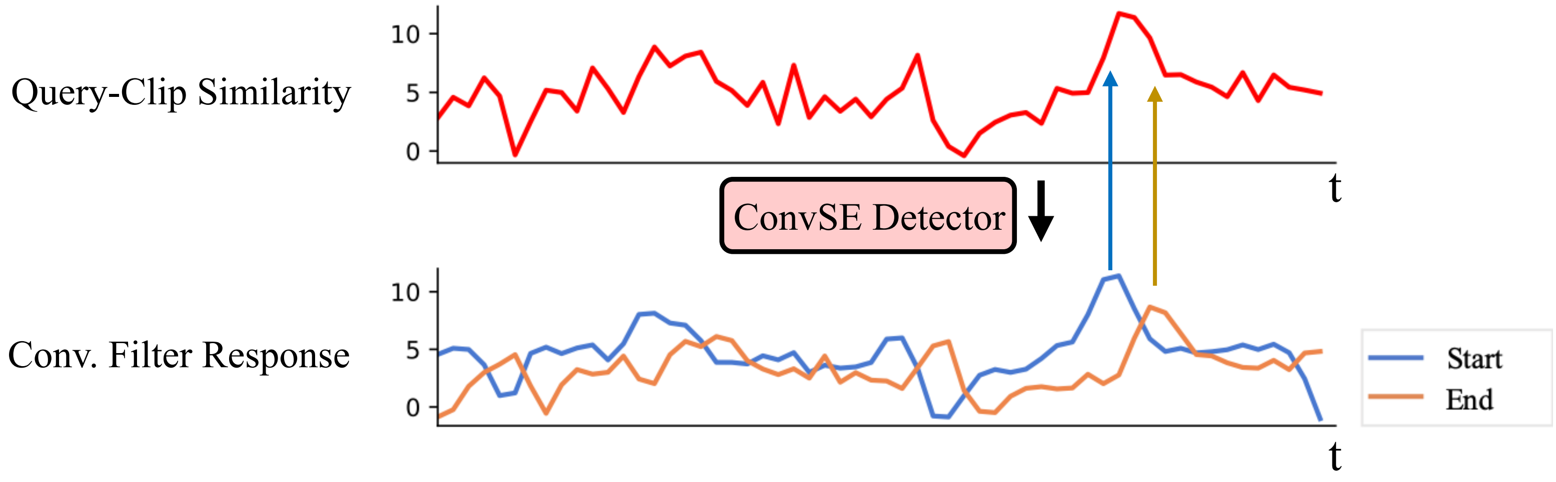
(Slides by Jie Lei)

Our Model: Cross-modal Moment Localization (XML)



(Slides by Jie Lei)

Our Model: Cross-modal Moment Localization (XML)



Single Video Moment Retrieval

- We first compute query-clip similarity scores $S_{\text{query-clip}} \in \mathbb{R}^l$.
- We then apply Convlutional Start-End (ConvSE) detector:

$$S_{\text{st}} = \text{Conv1D}_{\text{st}}(S_{\text{query-clip}}), \quad S_{\text{ed}} = \text{Conv1D}_{\text{ed}}(S_{\text{query-clip}})$$

- The scores are normalized with softmax to output the probabilities $P_{\text{st}}, P_{\text{ed}} \in \mathbb{R}^l$.

(Slides by Jie Lei)

Experiments

Baseline comparison on TVR test-public set, VCMR task.

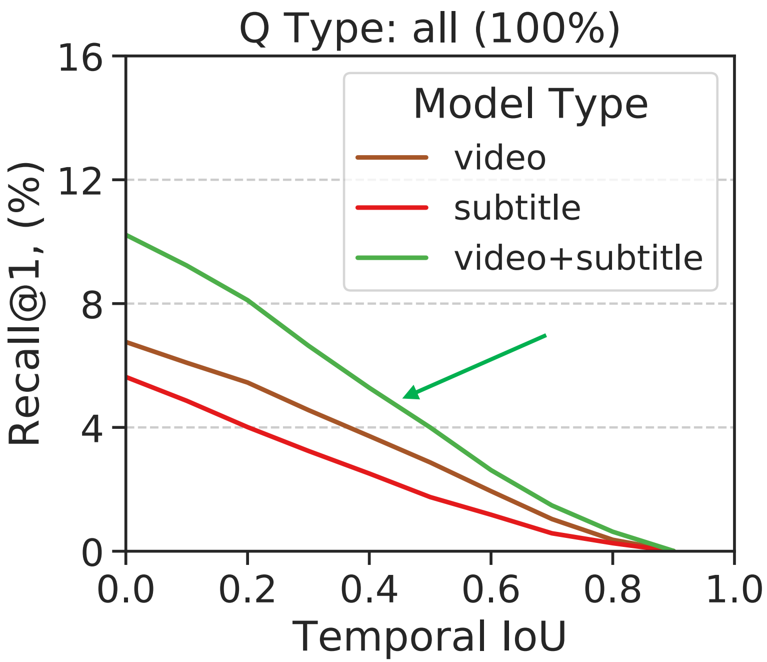
Model	w/ video	w/ sub.	IoU=0.5 R@1	IoU=0.5 R@5	IoU=0.5 R@10	IoU=0.5 R@100	IoU=0.7 R@1	IoU=0.7 R@5	IoU=0.7 R@10	IoU=0.7 R@100	Runtime ↓ (seconds)
Chance	-	-	0.00	0.02	0.04	0.33	0.00	0.00	0.00	0.07	
Proposal based Methods											
MCN	✓	✓	0.02	0.15	0.24	2.20	0.00	0.07	0.09	1.03	-
CAL	✓	✓	0.09	0.31	0.57	3.42	0.04	0.15	0.26	1.89	-
Retrieval + Re-ranking											
MEE+MCN	✓	✓	0.92	3.69	5.58	17.91	0.42	1.89	2.98	10.84	66.8
MEE+CAL	✓	✓	0.97	3.75	5.80	18.66	0.39	1.69	2.98	11.52	161.5
MEE+ExCL	✓	✓	0.92	2.53	3.60	6.01	0.33	1.19	1.73	2.87	1307.2
XML	✓	✓	**7.25**	**16.24**	**21.65**	**44.44**	**3.25**	**8.71**	**12.49**	**29.51**	**25.5**

(Slides by Jie Lei)

Experiments

Performance of XML models that use only video, subtitle, or both as inputs.

- Use both video and subtitle performs the best.

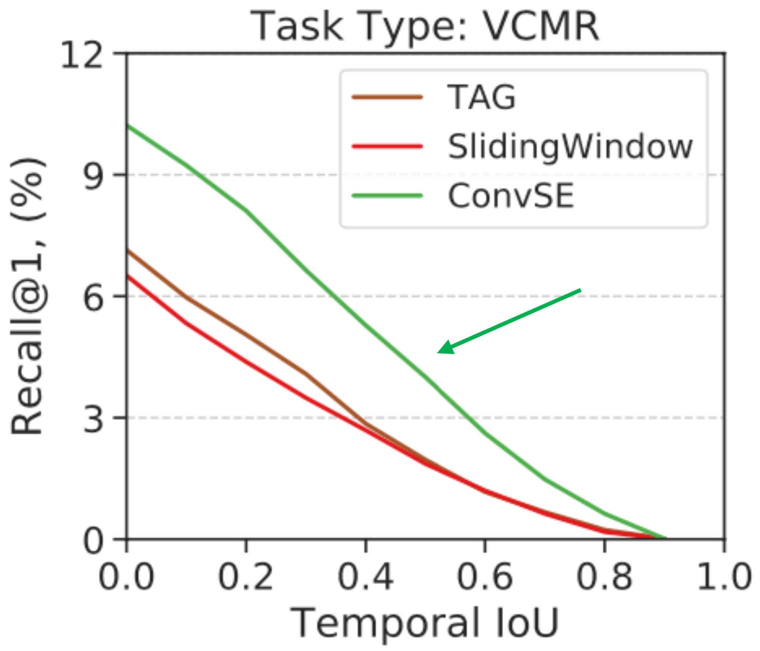


(Slides by Jie Lei)

Experiments

Performance comparison of moment generation methods, under the same XML backbone.

- TAG and SlidingWindow rely on handcrafted rules, while ConvSE learns from data.
- ConvSE performs consistently better across different IoU thresholds.



(Slides by Jie Lei)

Data & Code Release

TV show Retrieval (TVR):

TVR https://tvr.cs.unc.edu/

https://github.com/jayleicn/TVRetrieval

TVR Leaderboard

TVR tests a system's ability of localizing a moment from a large video (with subtitle) corpus. The performance is measured by R@K (Recall@K, K=1, 10, 100), with temporal IoU = 0.7.

Rank	Model	R@1	R@10	R@100
1 Jan 20, 2020	XML *UNC Chapel Hill* Paper, Code	3.32	13.41	30.52

TV show Captions (TVC):

-- We collected additional descriptions for each TVR moment.

TVC https://tvr.cs.unc.edu/tvc.html

https://github.com/jayleicn/TVCaption

TVC Leaderboard

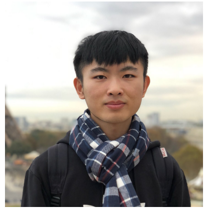
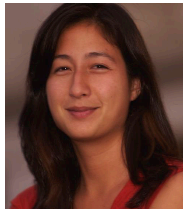
TVC requires systems to gather information from both video and subtitle to generate relevant descriptions. The performance is measured by B@4 (BLEU@4), M (METEOR), R (Rouge-L), C (CIDEr-D).

Rank	Model	B@4	M	R	C
1 Jan 20, 2020	MMT (video+sub) *UNC Chapel Hill* Paper, Code	10.87	16.91	32.81	45.38

(Slides by Jie Lei)

What is More Likely to Happen Next? Video-and-Language Future Event Prediction

UNC NLP

Jie Lei, Licheng Yu, Tamara L. Berg, Mohit Bansal
UNC Chapel Hill

(Slides by Jie Lei)

VLEP Task & Example

- **Task**: Given a video (with dialogue) as premise, predict what is most likely to happen next by selecting from two provided future events. This task requires using event schema knowledge, which is quite challenging for modern AI systems.

Premise Event

00:21,320 --> 00:23,381
[Mark] Oh yeah! Maybe a shake.

(*Premise Summary* : A woman with a white shirt with black buttons grinds fruit slush in a blender.)

Future Events (Which event is more likely to happen right after the premise?)

A. The woman in the white shirt pours the slush into a cup.
(*Rationale*: Slushy drinks are more commonly served in a cup, but there are hollowed out watermelon rinds sitting around the blender.)

B. The woman in the white shirt pours the slush into a watermelon rind and passes it to Mark.
(*Rationale*: There are hollowed out watermelon rinds sitting around the blender.)

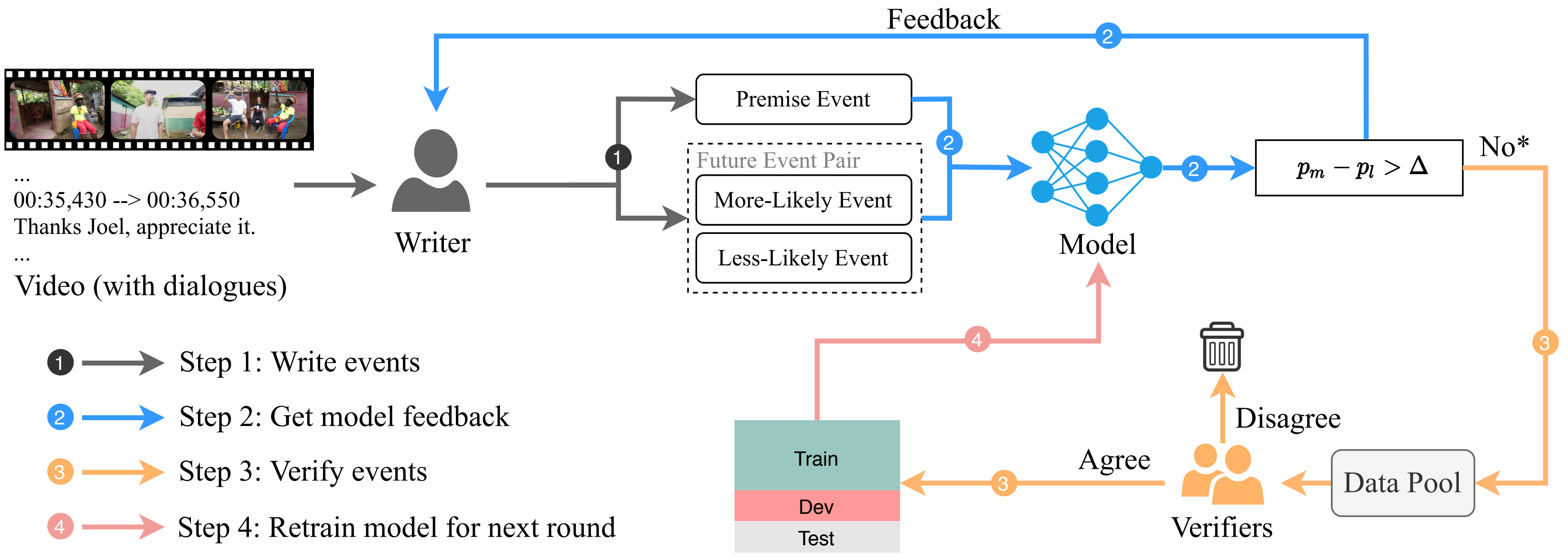
A VLEP example with a YouTube Vlog video.

(Slides by Jie Lei)

VLEP Dataset Collection

- Human-and-Model-in-the-Loop Adversarial Data Collection.

- Adversarial Matching.
  - We sample negatives from existing human positives that is close to the given premise but not overly similar to the true positive.



Adversarial collection procedure.

(Slides by Jie Lei)

[Nie et al., ACL 2020 , Zellers et al., CVPR 2019]

VLEP Dataset

- We collected 28.7K examples with 10K TV show and YouTube Vlog video clips from different genres. We also show top unique verbs in each genre.

Domain	Genre	#Shows (#Channels)	#Videos	#Examples
TV show	Sitcom	3	4,117	12,248
	Medical	2	1,558	5,198
	Crime	1	1,072	4,306
YouTube Vlogs	Travel, Food	6	2,406	4,812
	Family, Daily	3	1,081	2,162
Total	-	15	10,234	28,726

Data statistics by genre.

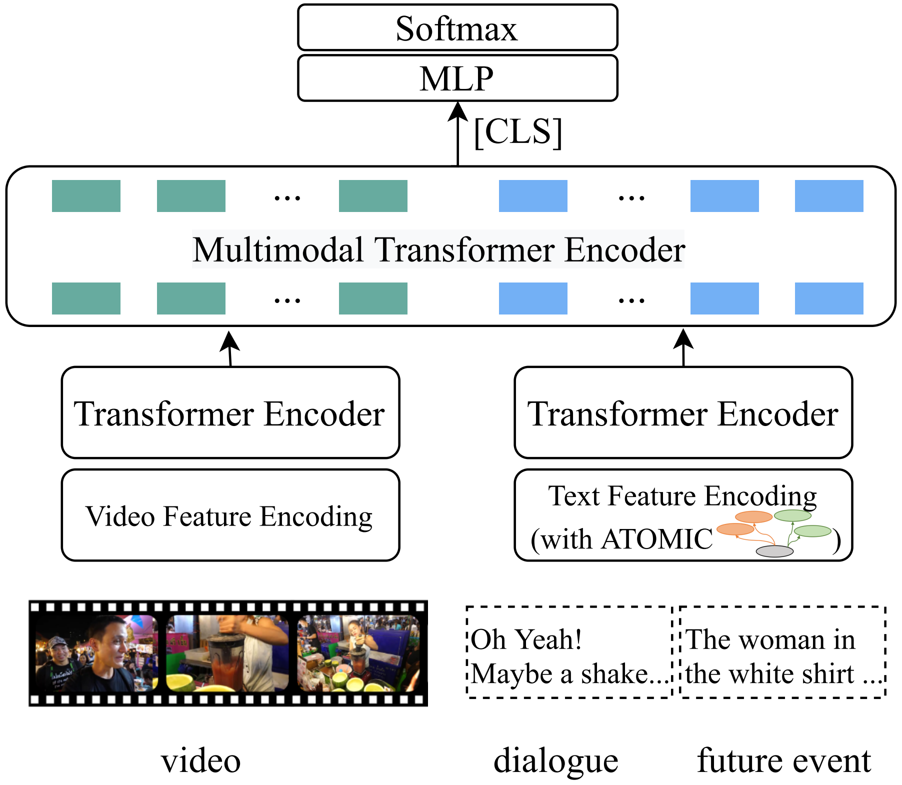
Genre	Top Unique Verbs
Sitcom	change, offer, hear, should, accept, yell, hang, join, apologize, shut, shout, realize
Medical	die, treat, cry, yell, smile, proceed, examine, approach, argue, save, admit, rush
Crime	kill, shoot, point, question, toss, hang, remove, catch, lie, deny, investigate,
Travel, Food	taste, add, pour, dip, cook, describe, cut, order, serve, stir, prepare, enjoy, buy
Family, Daily	drive, jump, wear, point, smile, touch, climb, dress, set, swim, hide, lay, blow

Top unique verbs in each genre.

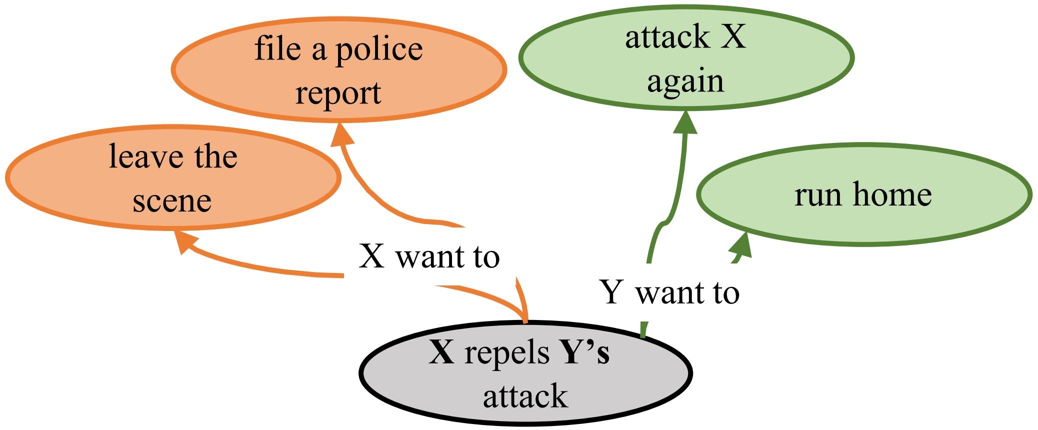
(Slides by Jie Lei)

Method

- A transformer-based method.
  - Video feature: 2D appearance feature + 3D action feature.
  - Text feature: from a RoBERTa model fine-tuned on event schemas from ATOMIC knowledge base.
  - A multimodal transformer encoder for both video and text.



Model overview



An ATOMIC example. Example sentence: X repels Y's attack, then X want to file a police report.

(Slides by Jie Lei)

[Sap et al., AAAI 2019]

Experiments

- We split data into 70% training, 15% development, 15% testing splits.
- Left: Video, dialogue are both useful for the task, when combined, we obtain the best performance of 67.46%, but is still far below human performance of 90.5%.
- Right: event schema knowledge is useful for the task, without ATOMIC sentences for fine-tuning, we see a lower performance.

Model	Accuracy (%)
chance	50.00
future only	58.09
video + future	59.03
dialogue + future	66.63
video + dialogue + future	67.46
human (dialogue + future)	76.25
human (video + dialogue + future)	90.50

Results on VLEP test splits.

Model	Accuracy (%)
video + dialogue + future	67.46
- ATOMIC fine-tuning	66.96

Effect of ATOMIC fine-tuning.

(Slides by Jie Lei)

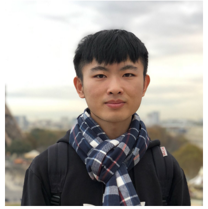
Data Release

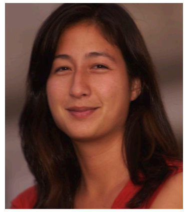
https://github.com/jayleicn/VideoLanguageFuturePred

(Slides by Jie Lei)

Thanks!

UNC NLP






Jie Lei, Licheng Yu, Tamara L. Berg, Mohit Bansal
UNC Chapel Hill

(Slides by Jie Lei)