COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Multimodal Knowledge Graphs: Automatic Extraction & Applications

Prof. Shih-Fu Chang

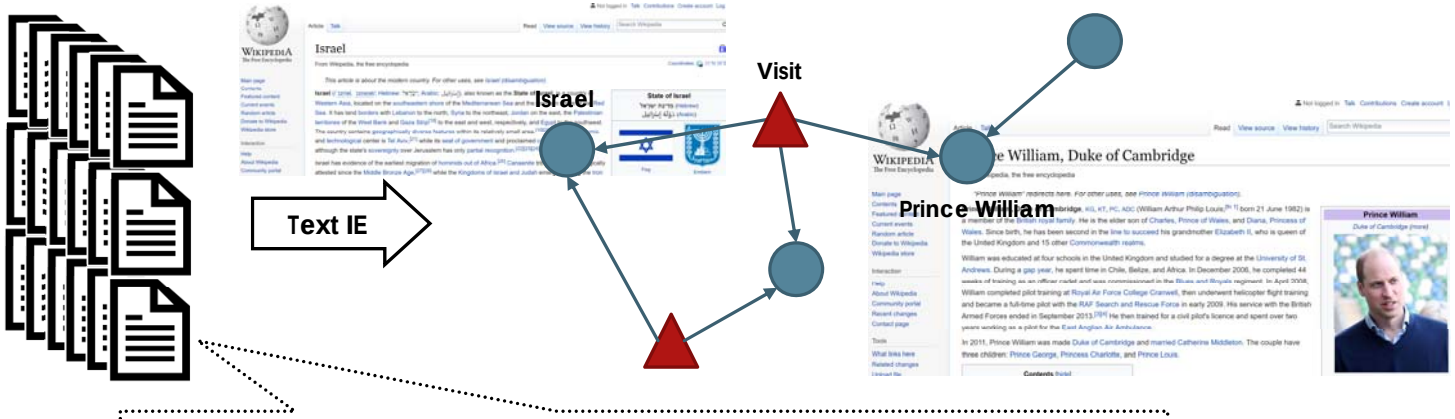
In Collaboration with
Alireza Zareian, Hassan Akbari, Brian Chen
Columbia University

Prof. Heng Ji,
Spencer Whitehead, Manling Li
RPI/UIUC

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Knowledge Graphs

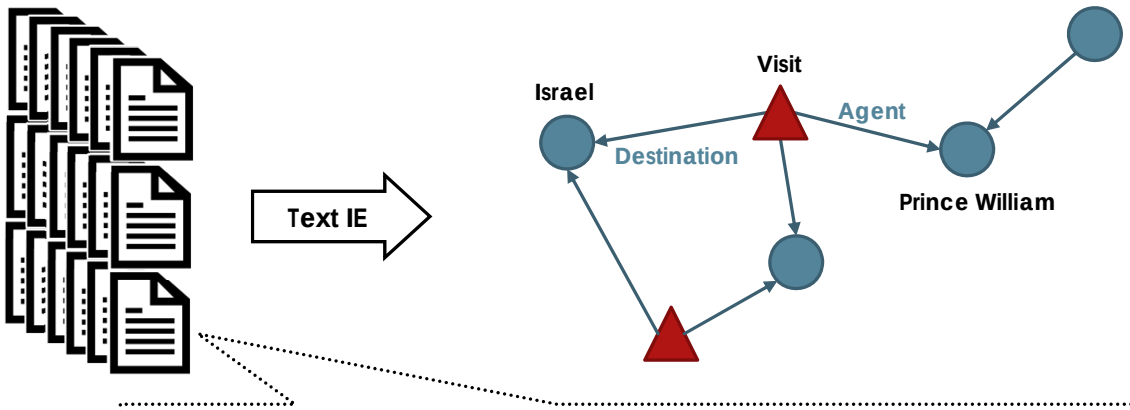
- NLP: Information extraction from text
  - Entities, events, relations, etc.



The first-ever official visit by a British royal to Israel is underway. Prince William the 36-year-old Duke of Cambridge and second in line to the throne will meet with both Israeli and Palestinian leaders over the next three days.

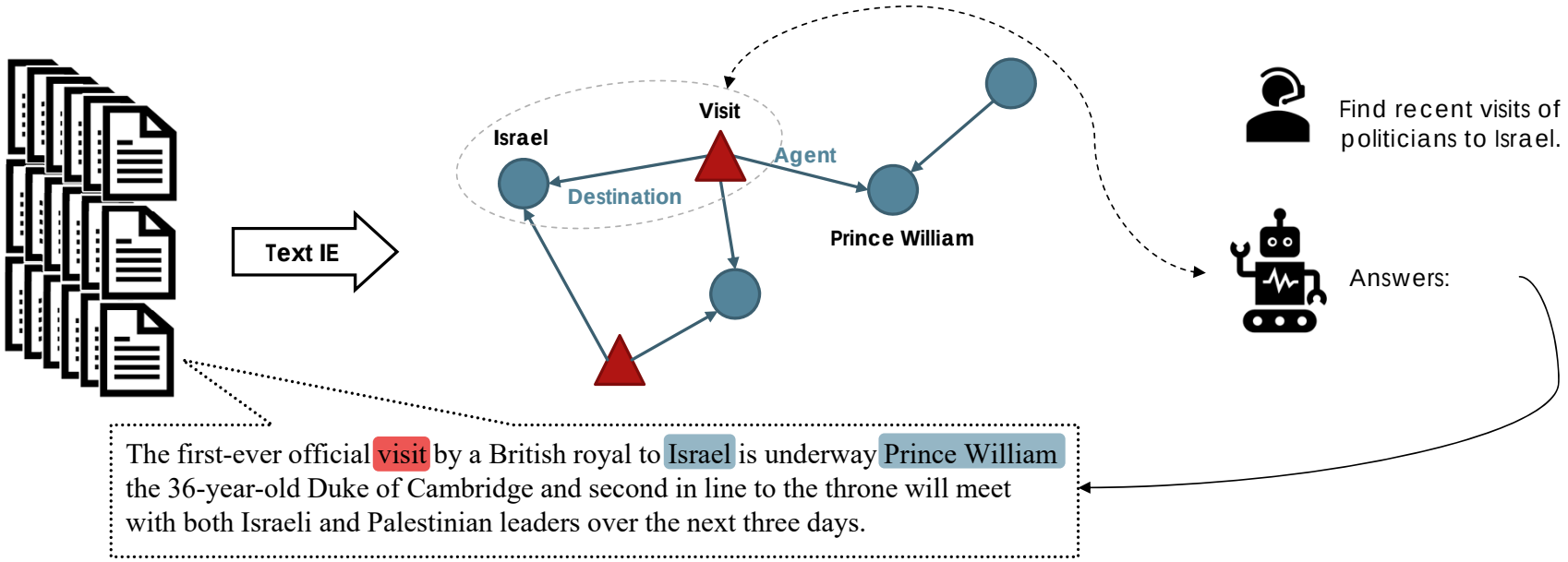
Knowledge Graphs

- NLP: Information extraction from text
  - Entities, events, relations, etc.
  - Event-centric, Describe What Happens
    - Entities are characterized by the argument *role* they play in events



Knowledge Graphs

- NLP: Information extraction from text
- Application: Question Answering

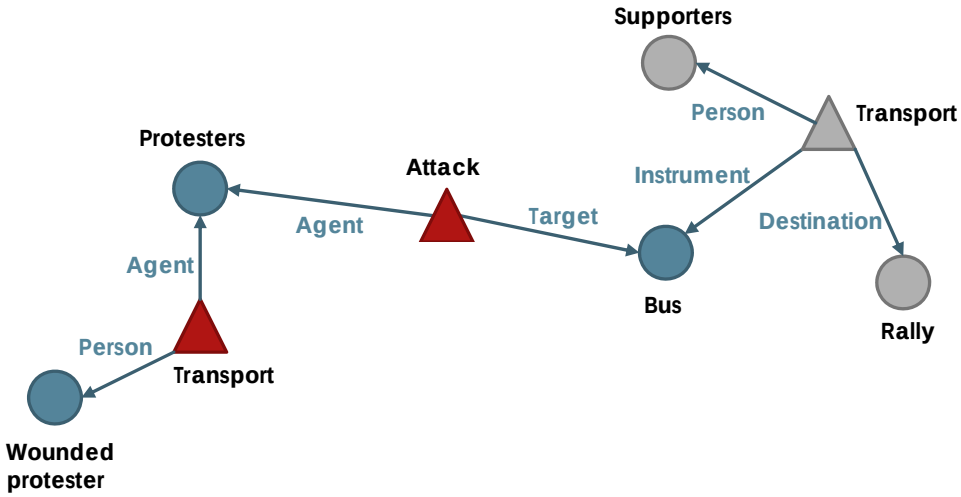


Why Multimodal?

- Visual data contains complementary data that can be used for:
  - Visual Illustration
  - Disambiguation
  - Additional Details

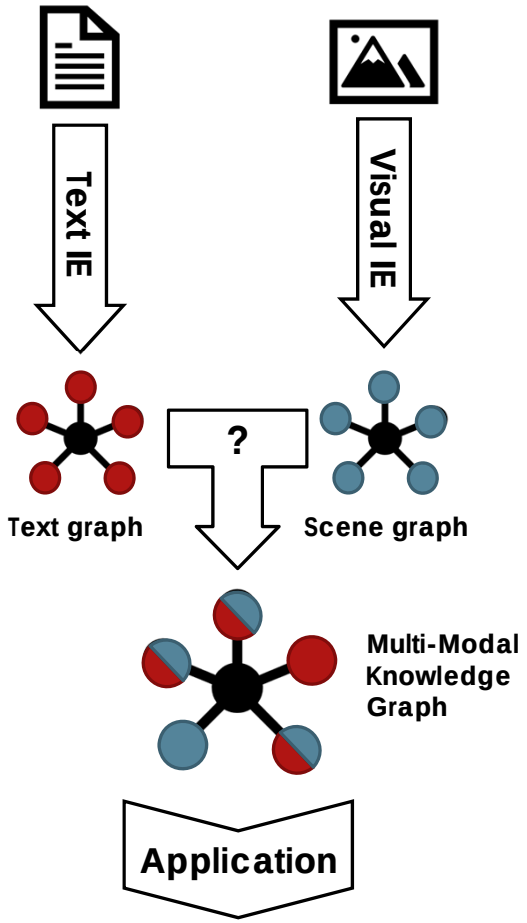
News Article: Thai opposition protesters[Attacker] attack[Attack] a bus[Target] carrying pro-government Red Shirt supporters on their way to a rally. Protesters[Agent] are carrying [TransportPerson] a wounded protester[Person] to . ...

Multimodal KG



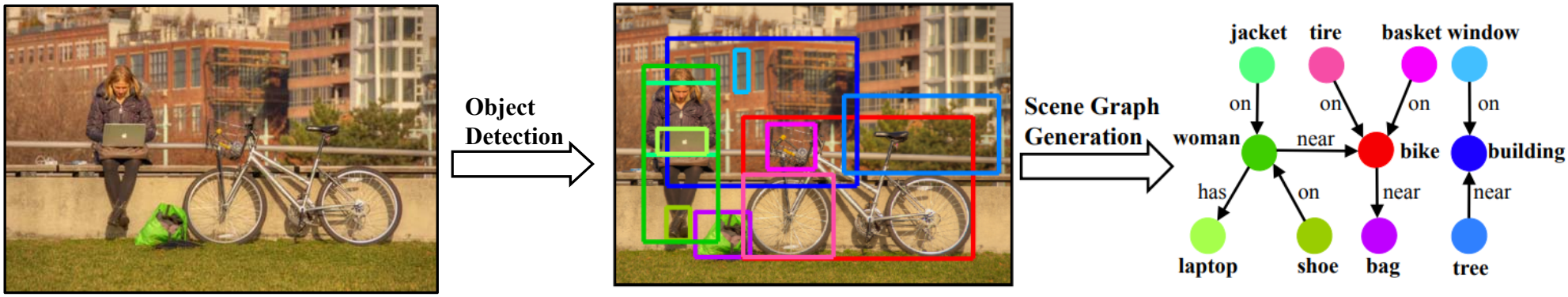
Challenges & Applications

- Challenges:
  - Parsing text to structured semantic graph
  - Parsing images/videos to structures
  - Grounding event/entities across modalities
  - Multimodal argument role
- Applications
  - Story Generation and Summarization
  - Question Answering
  - Commonsense Discovery



Challenge 1: Parsing Images to Scene Graphs

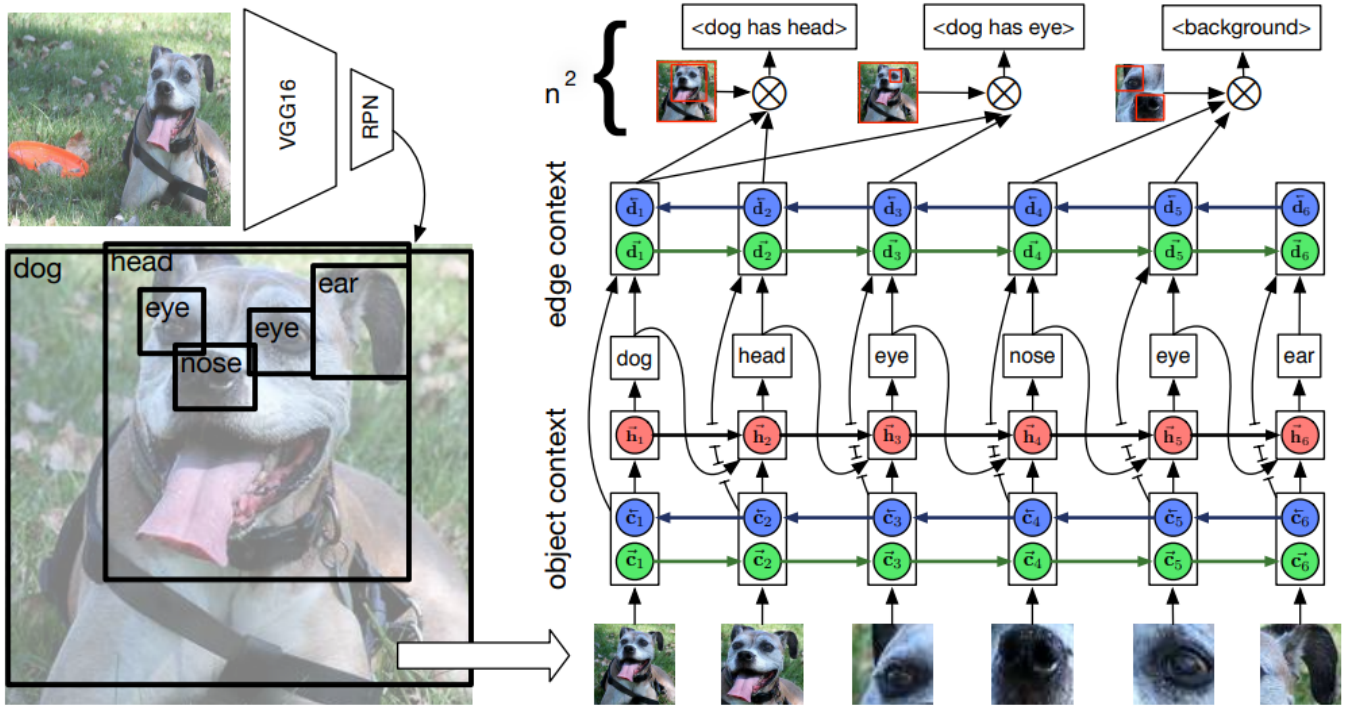
- Extract structured representation of a scene
  - Entities and their semantic relationships



- Applications
  - Reasoning, Q&A, common sense discovery
- Challenges
  - Annotation cost, computational complexity, limited fixed vocabulary

Parsing Images to Scene Graphs

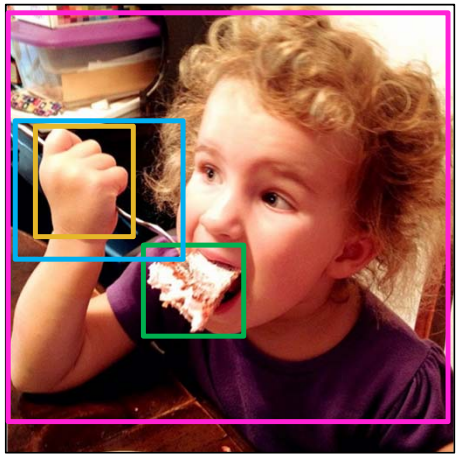
- Existing methods
  - Extract region proposals
  - Contextualize features by RNN (or message passing)
  - Classify all nodes and pairs of nodes
- Limitations
  - Computationally exhaustive
    - $O(n^2)$ for $n \approx 300$ proposals
  - Difficult to model higher order relationships, e.g. *"girl eating cake with fork"*
  - Requires full supervision

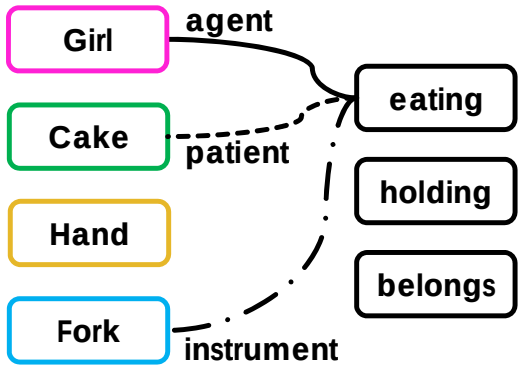


Neural Motifs (Zellers, Yatskar, Thomson, Choi, CVPR 2018)
One of the SOTA methods for scene graph generation

Reformulate as an Event-Centric Problem

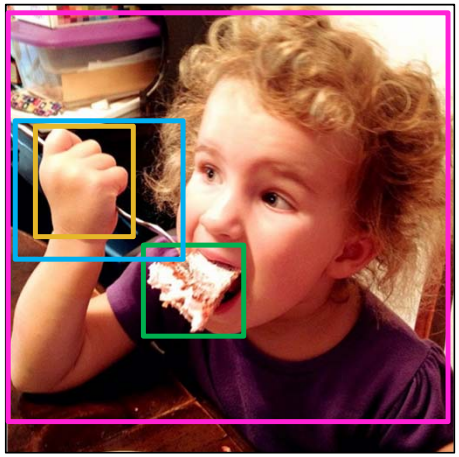
- Our work: Visual Semantic Parsing Network (Zareian et al. 2019)
  - Generalized formulation of scene graph generation
    - Entity-centric → balanced representation of predicates & entities
    - Model argument role relations beyond (subject, object), (agent, patient) relations
    - Reduce computational complexity from $O(n^2)$ to sub-quadratic

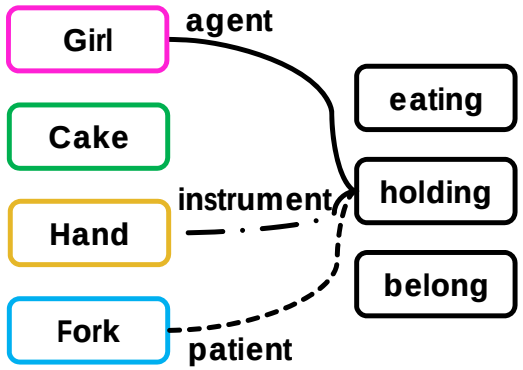




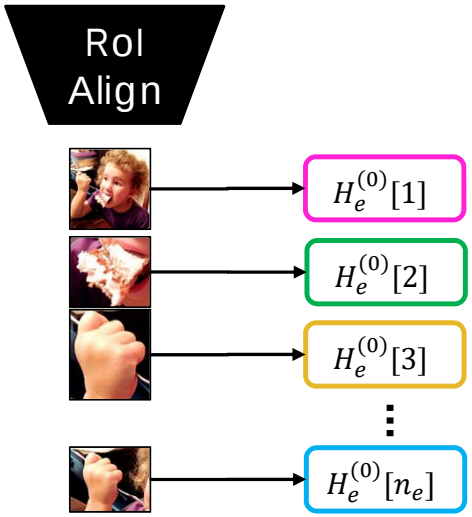
Reformulate as an Event-Centric Problem

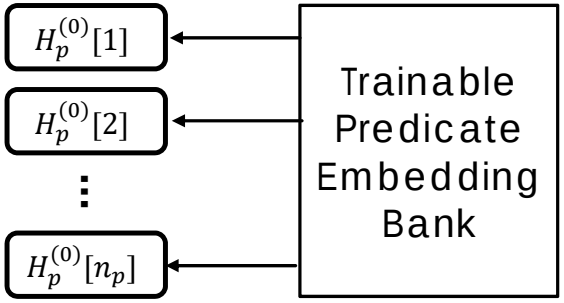
- Our work: Visual Semantic Parsing Network (Zareian et al. 2019)
  - Generalized formulation of scene graph generation
    - Entity-centric → balanced representation of predicates & entities
    - Model argument role relations beyond (subject, object), (agent, patient) relations
    - Reduce computational complexity from $O(n^2)$ to sub-quadratic





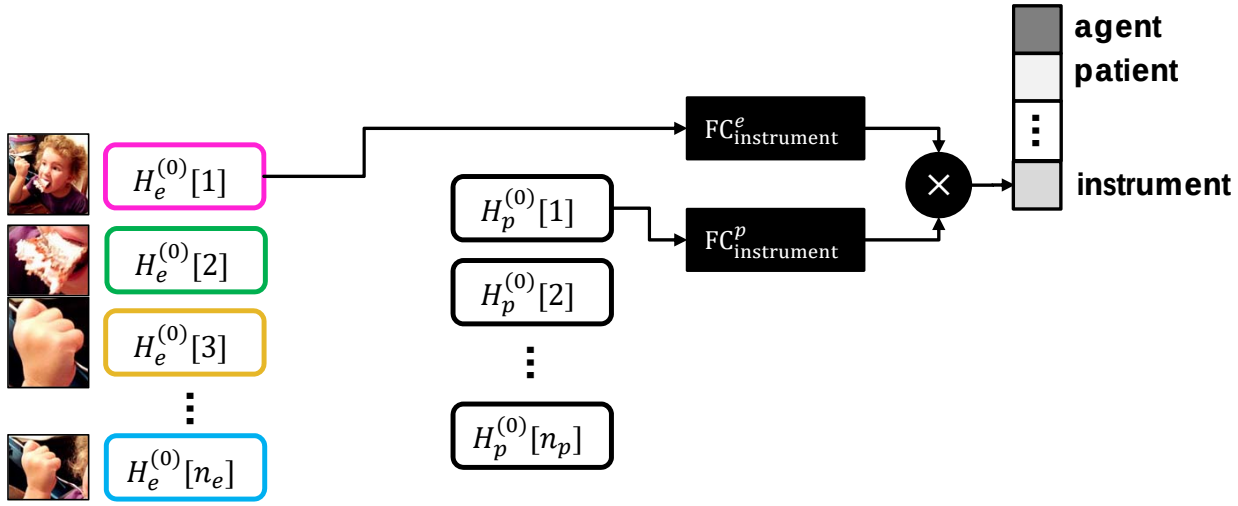
Bi-Partite Embeddings for Entity & Predicate





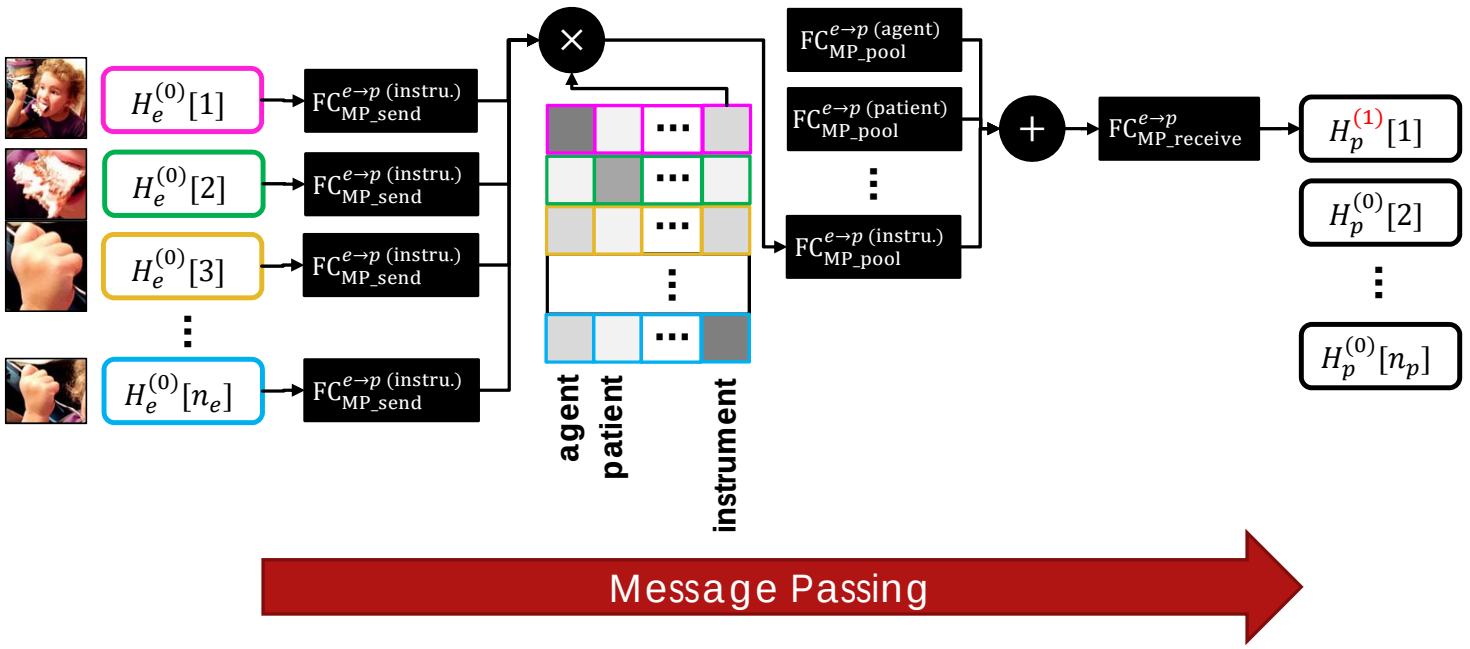
Argument Role Prediction

- Initialize entity and predicate nodes
- Compute role-specific attention scores
  - Input: entity-predicate feature pairs
  - Output: scalar for each thematic role



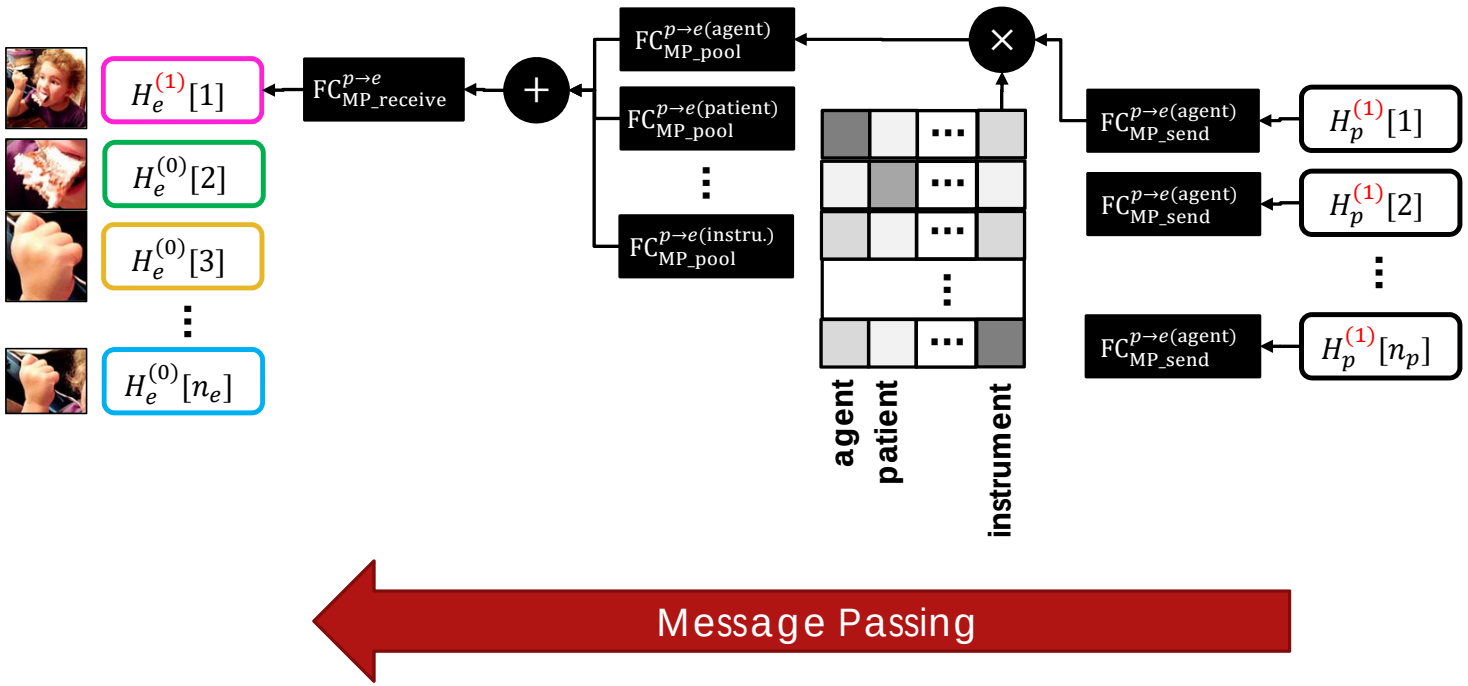
Role-Dependent Message Passing

- **Bi-directional Message passing**
- **Entities → Roles → Predicates**



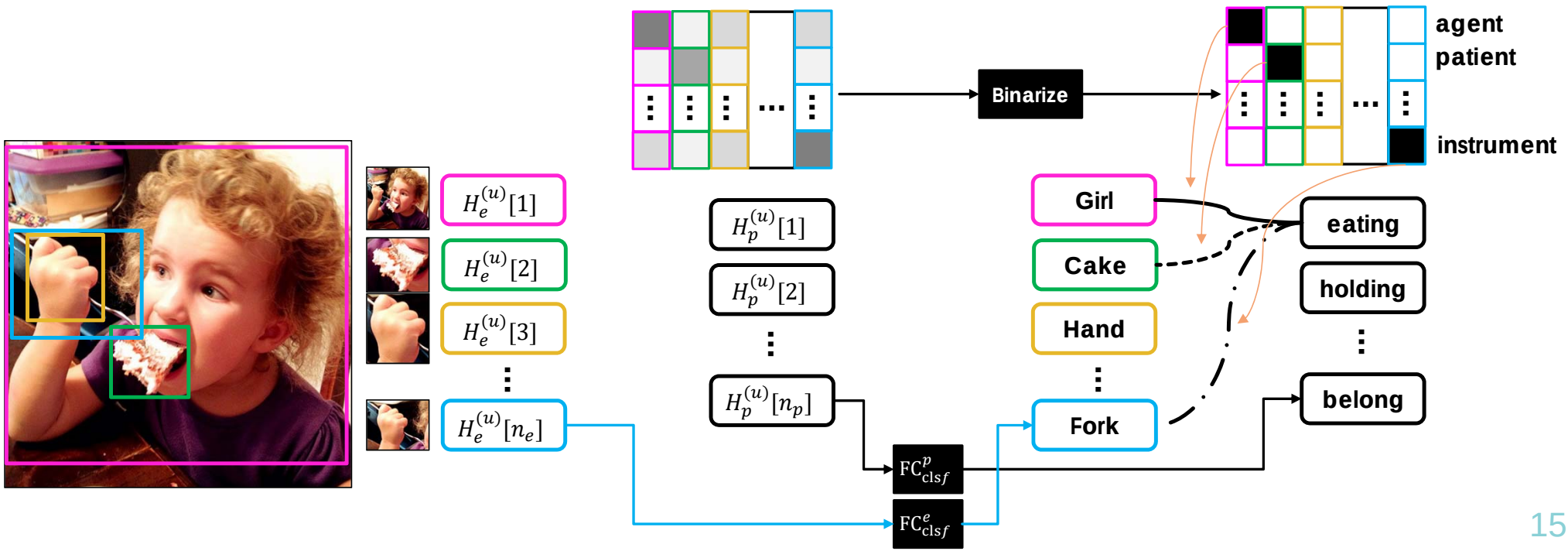
Role-Dependent Message Passing

- **Bi-directional Message passing**
- **Entities ← Roles ← Predicates**



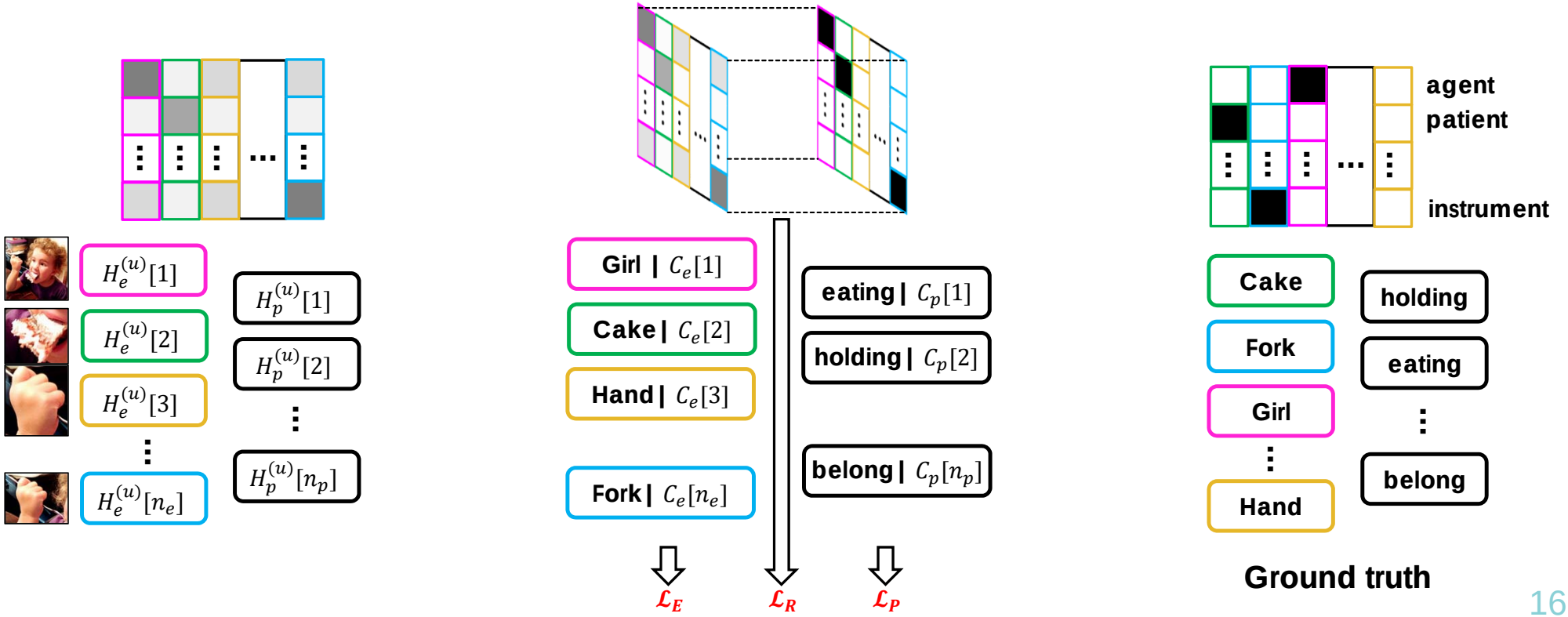
Visual Semantic Parsing Network

- Bi-directional Message passing
- Repeat for u iterations
- Classify nodes and edges



Visual Semantic Parsing Network

- **Weakly supervised training**
  - Unknown alignment between output and ground truth graphs

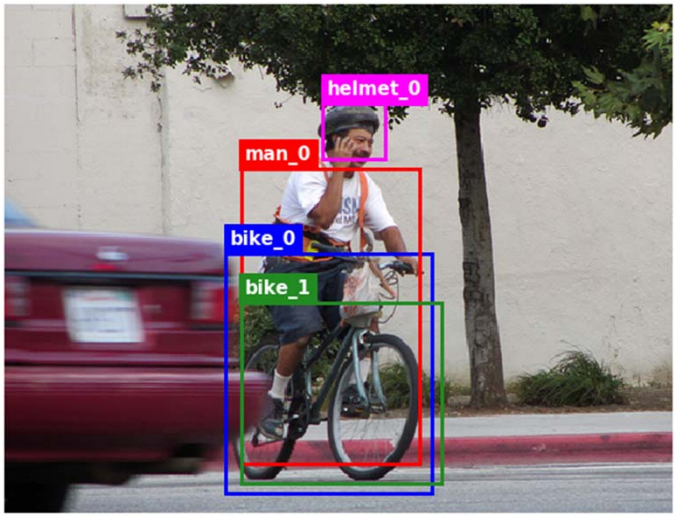


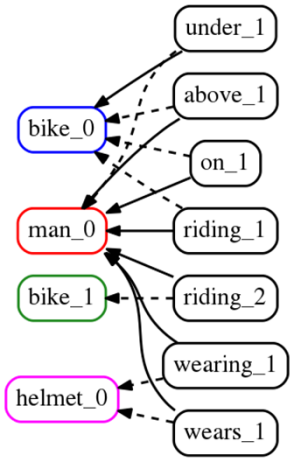
Visual Semantic Parsing Network

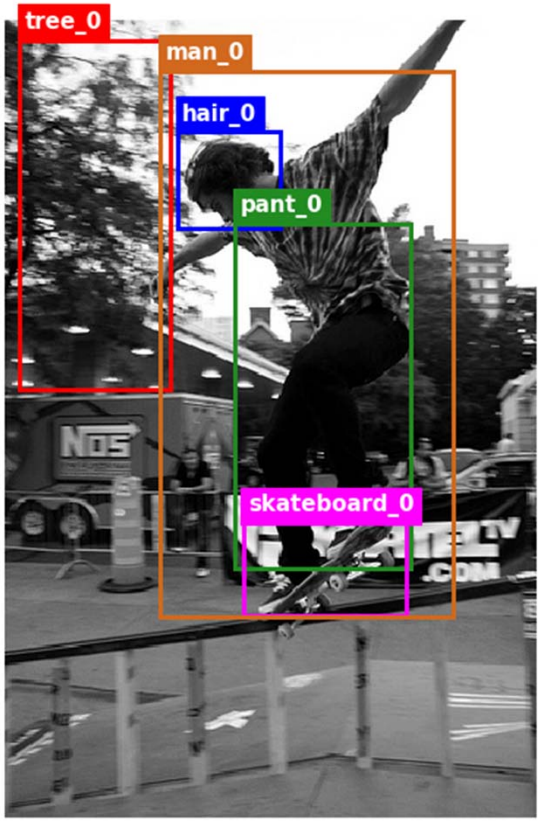
- Accuracy: major improvement for weakly supervised
- Speed: 10X-20X faster

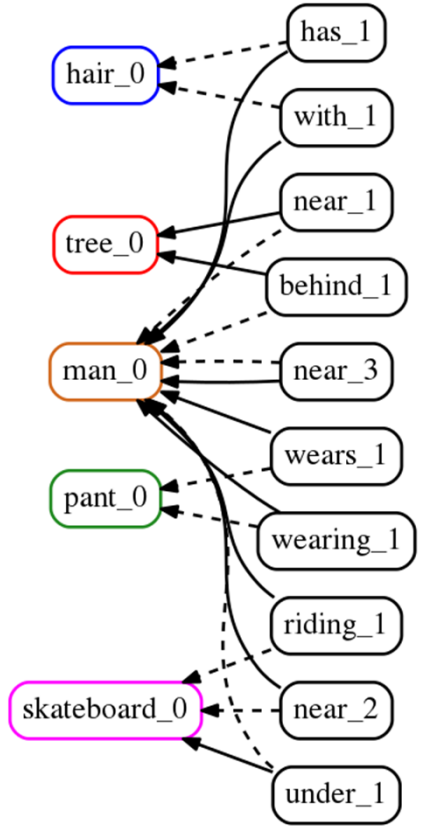
Method	Supervision	SGGen		SGCls		PredCls		Time
		R@50	R@100	R@50	R@100	R@50	R@100	
IMP	Full	3.4	4.2	21.7	24.4	44.7	53.1	1.64
MSDN	Full	7.7	10.5	19.3	21.8	63.1	66.4	3.56
MotifNet	Full	6.9	9.1	23.8	27.2	41.8	48.8	N/A
Assoc. Emb.	Full	9.7	11.3	26.5	30.0	**68.0**	**76.2**	N/A
Graph R-CNN	Full	**11.4**	**13.7**	**29.6**	**31.6**	54.2	59.1	N/A
Factorizable Net	Full	N/A	N/A	N/A	N/A	N/A	N/A	0.55
ViSParNet (Ours w/ VGG)	Full	4.0	4.8	20.8	25.4	39.6	56.1	**0.14**
ViSParNet (Ours)	Full	5.5	6.4	22.4	27.2	39.7	55.6	**0.11**
VtransE-MIL	Weak	0.7	0.9	1.5	2.0	N/A	N/A	
PPR-FCN	Weak	1.5	1.9	2.4	3.2	N/A	N/A	
ViSParNet (Ours w/ VGG)	Weak	1.3	1.5	10.3	12.0	41.2	55.1	
ViSParNet (Ours)	Weak	**2.6**	**3.0**	**12.9**	**15.2**	**42.7**	**58.3**	

Visual Semantic Parsing Network



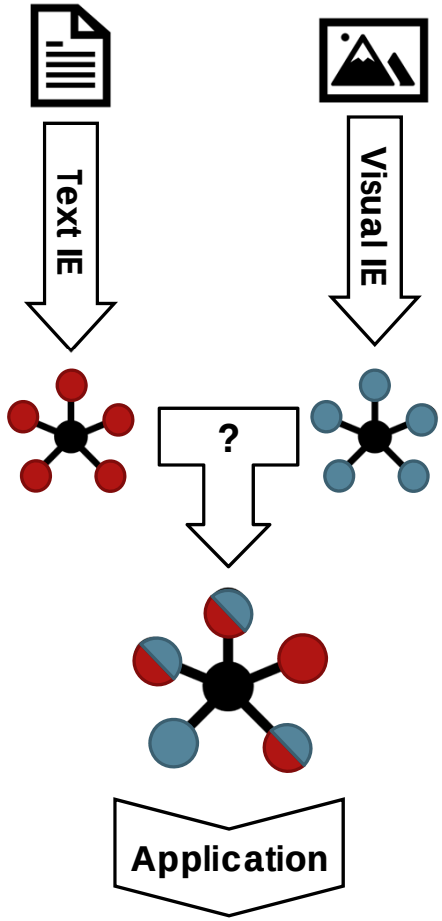






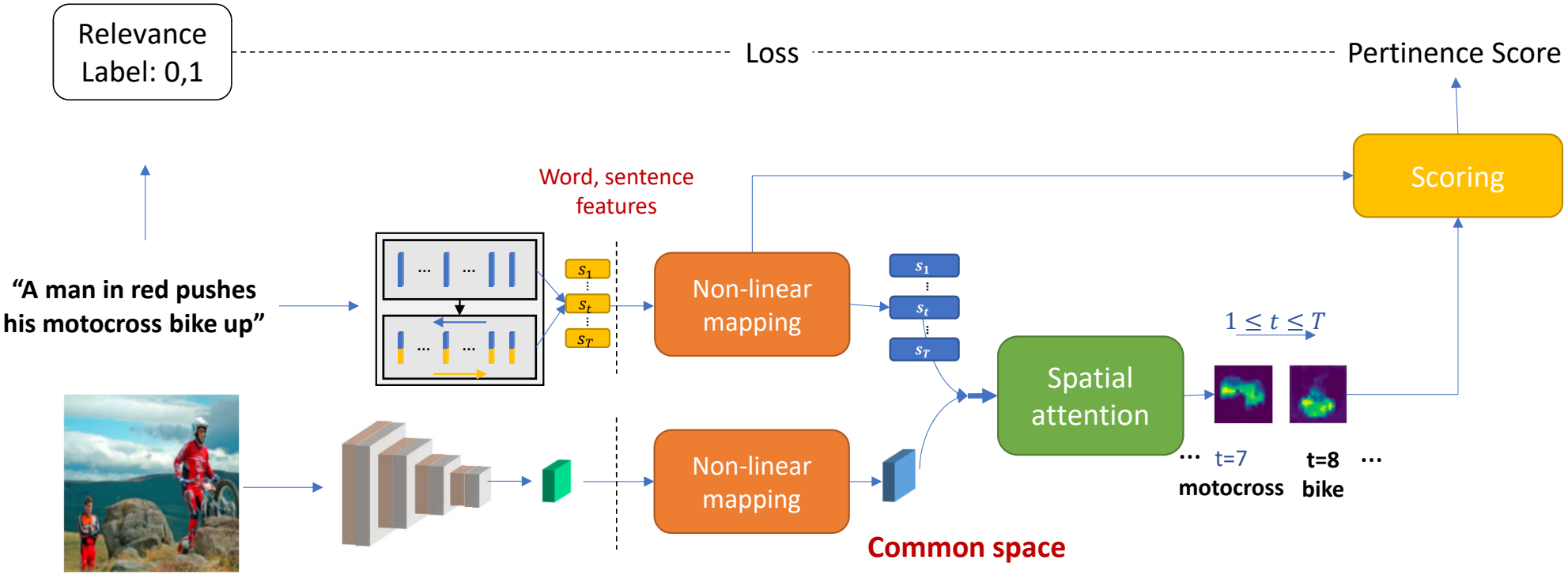
Challenge 2: Visual Grounding

- Localize text query in image
  - Bridge visual and text knowledge graphs
  - Without using predefined classifiers



- Challenges
  - Annotation Cost
  - Limited training data (domain specific)
  - Abstract concept not groundable

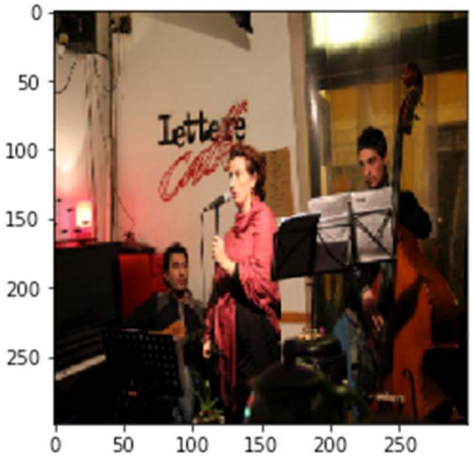
Typical Approach to Visual Grounding



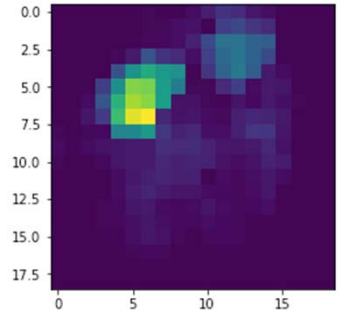
Heatmap Visualization

$H_l \in \mathbb{R}^{T \times N}$ → reshaped to original feature map size $H \in \mathbb{R}^{T \times h \times w}$

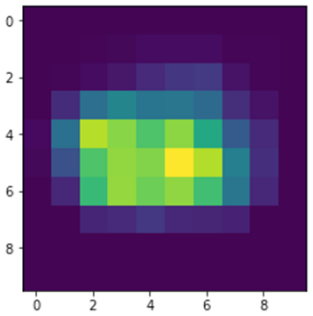
Each word is assigned with a heatmap
(One heatmap for an entire sentence)



"Two men play a cello and a guitar as a woman sings at the Lettere Caffe."

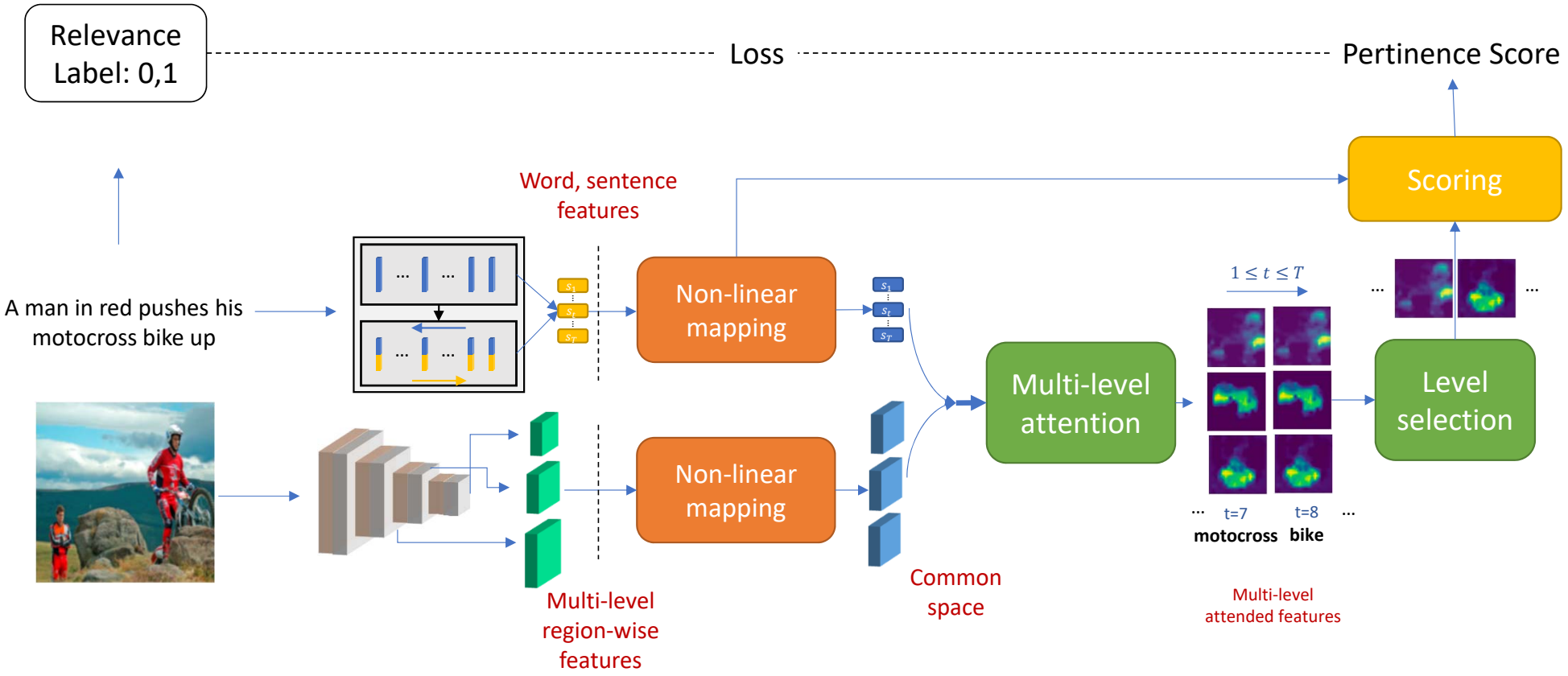


Heatmap for "*Lettere Caffe*"



Heatmap for sentence

Multi-level Attention Model for Grounding



Akbari, H., S. Karaman, S. Bhargava, B. Chen, C. Vondrick, and S.-F. Chang. "Multi-level Multimodal Common Semantic Space for Image-Phrase Grounding." CVPR 2019.

Multi-Level Attention Helps

- Multi-level attention weights selected for words in different semantic categories

Level / PNASNet Layers	Selection Rate (%)									
	bodyparts	animals	people	instrument	vehicles	scene	other	clothing	average	sentence
1 / Cell 5	2.6	10.4	7.5	0.9	2.0	5.4	5.4	5.3	6.3	0.7
2 / Cell 7	0.1	2.0	4.2	0.0	1.7	2.5	0.9	0.3	2.5	0.05
3 / Cell 9	85.9	48.4	64.6	88.6	68.3	49.5	70.9	86.1	66.5	86.51
4 / Cell 11	11.4	39.2	23.7	10.5	27.9	42.6	22.8	8.3	24.7	12.7

Table 3. Level selection rate for different layers of PNASNet on different categories in Flickr30k

Challenge 2: Visual Grounding

- Results

			Test Accuracy		
Method	Settings	Training	VG	Flickr30k	ReferIt
Baseline	Random	-	11.15	27.24	24.30
Baseline	Center	-	20.55	49.20	30.40
TD [59]	Inception-2	VG	19.31	42.40	31.97
SSS [17]	VGG	VG	30.03	49.10	39.98
Ours	BiLSTM+VGG	VG	50.18	57.91	**62.76**
Ours	ELMo+VGG	VG	48.76	60.08	60.01
Ours	ELMo+PNASNet	VG	**55.16**	**67.60**	61.89
CGVS [41]	Inception-3	MSR-VTT	-	50.10	-
FCVC [10]	VGG	MSCOCO	14.03	29.03	33.52
VGLS [47]	VGG	MSCOCO	24.40	-	-
Ours	BiLSTM+VGG	MSCOCO	46.99	53.29	47.89
Ours	ELMo+VGG	MSCOCO	47.94	61.66	47.52
Ours	ELMo+PNASNet	MSCOCO	**52.33**	**69.19**	**48.42**

Table 1. Phrase localization accuracy (pointing game) on Flickr30k, ReferIt and VisualGenome (VG) compared to state of the art methods.

Examples of Visual Grounding

- Examples

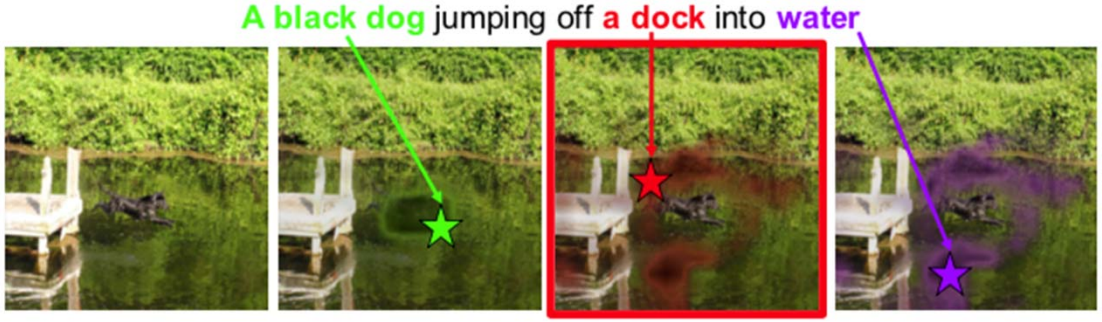
Figure 5. Image-sentence pair from Flickr30k with four queries (colored text) and corresponding heatmaps and selected max value (stars).

Figure 2. Some image-sentence pairs from Flickr30K, with two queries (colored text) and corresponding heatmaps and selected max value (stars).

Examples of Visual Grounding

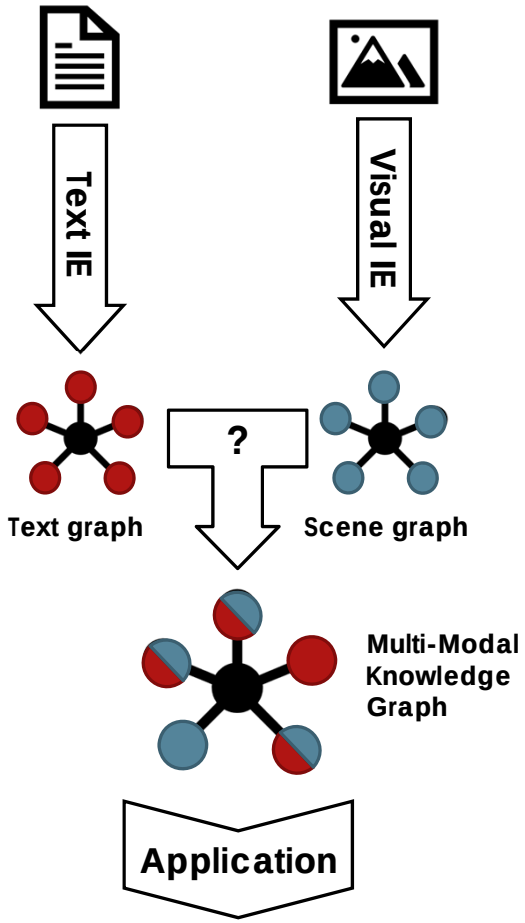
- Examples

Figure 7. Some failure cases of our model. The model makes some semantically reasonable mistakes in pointing to regions.



Multimodal KG Challenges & Applications

- Challenges:
  - Parsing text to structured semantic graph
  - Parsing images/videos to structures
  - Grounding entities across modalities
  - Multimodal argument role
- Applications
  - Story Generation and Summarization
  - Question Answering
  - Commonsense Discovery



Challenge 3: Multimodal Argument Role

- Event argument role labeling in NLP

- How to use image to extract unmentioned information about an event?
  - e.g. location, participants, instruments, etc.

Related Work

- Prior Work: Situation Recognition
  - Given an image:
    - Classify the event
    - For each argument role type, classify entity type
      - E.g. the attacker is a person
- Multimodal Argument Role
  - Assign images to dynamic event mentions in text rather than predefined event types
  - Recognize argument roles played
  - Localize argument roles

Spraying	
ROLE	**VALUE**
AGENT	MAN
SOURCE	SPRAY CAN
SUBSTANCE	PAINT
DESTINATION	WALL
PLACE	ALLEYWAY

Spraying	
ROLE	**VALUE**
AGENT	FIREMAN
SOURCE	HOSE
SUBSTANCE	WATER
DESTINATION	FIRE
PLACE	OUTSIDE

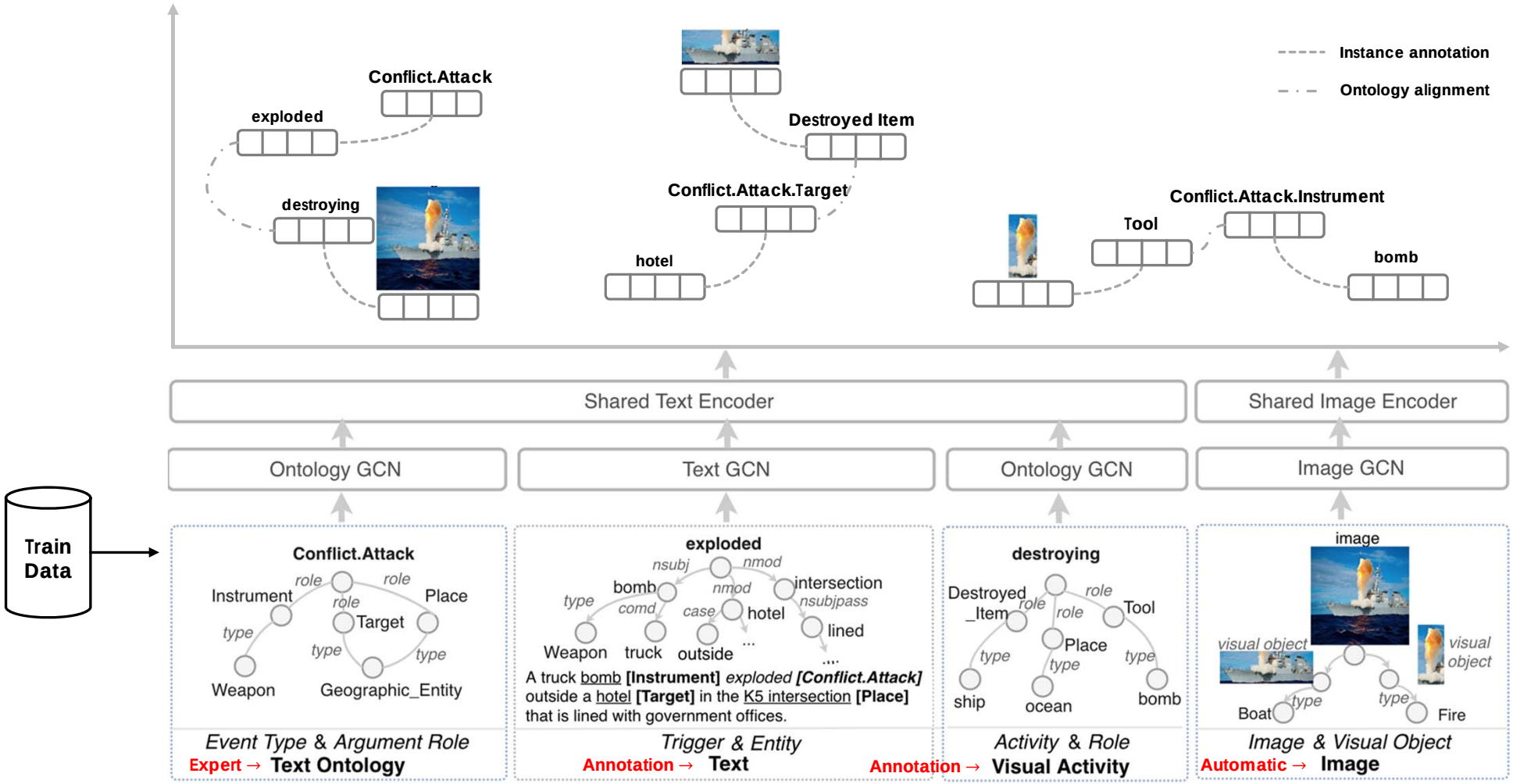
Yatskar, Zettlemoyer, and Farhadi, ECCV 2016

Multimodal Argument Role: Task

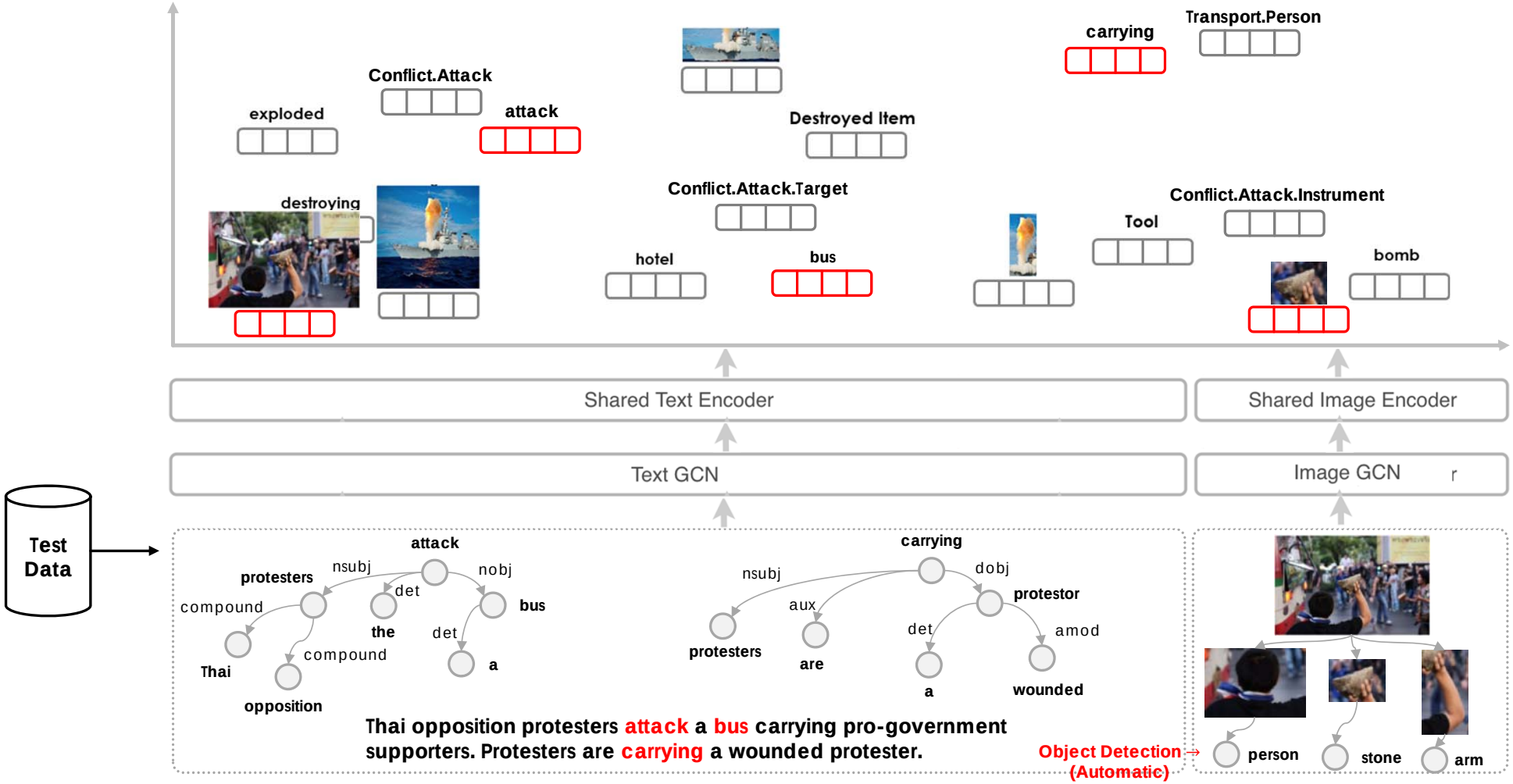
- Input: a text document
  - Main text
  - Images
  - Captions (optional)
- Output:
  - Multiple event types from text
  - Argument role labels from text
  - Assign each image to one of the mentioned events
  - Discover and localized argument role labels on image

News Article: Thai opposition protesters[Attacker] attack[Attack] a bus[Target] carrying pro-government Red Shirt supporters on their way to a rally. Protesters[Agent] are carrying [TransportPerson] a wounded protester[Person] to . ...

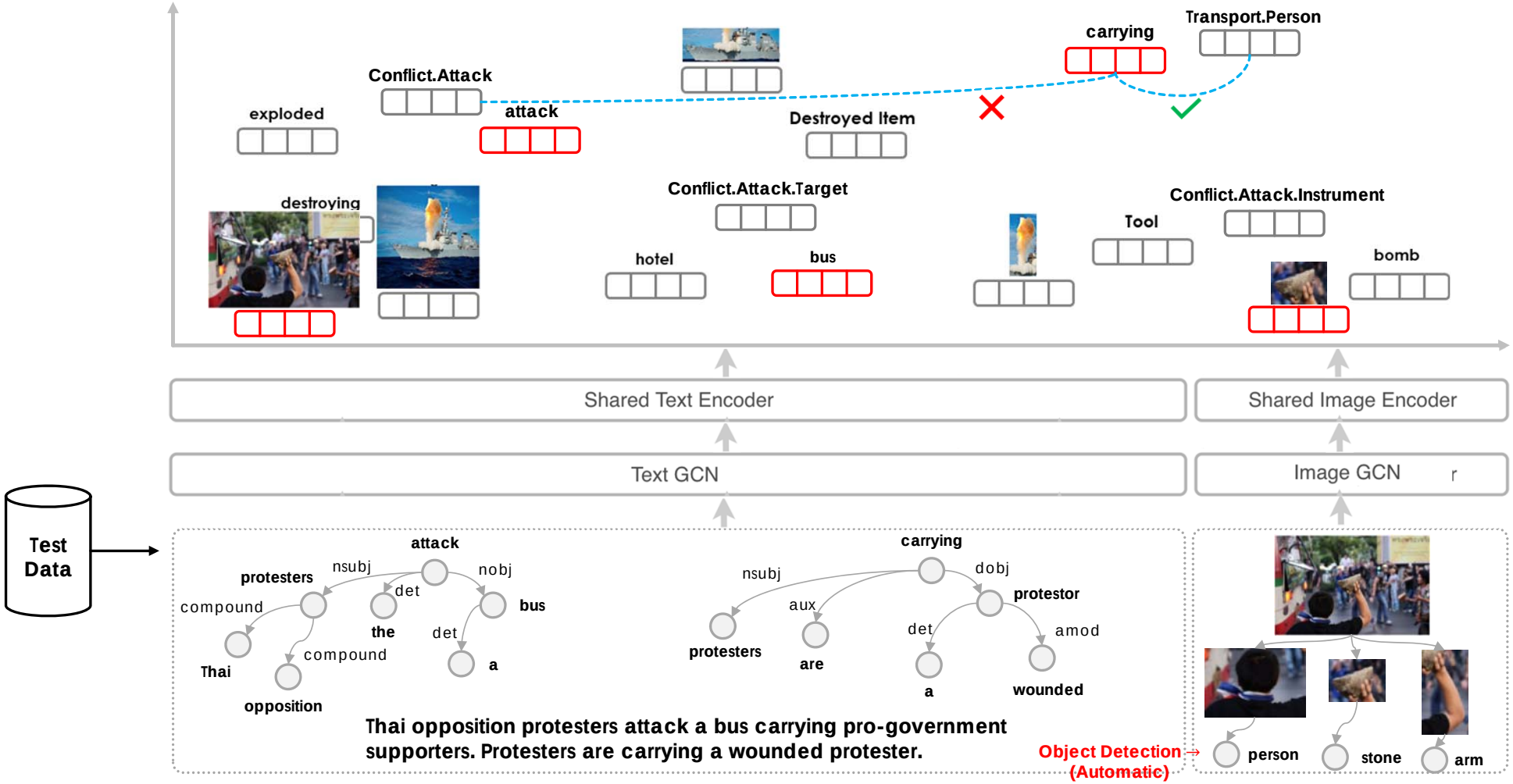
Idea: Multimodal Common Representation for Ontology & Instances



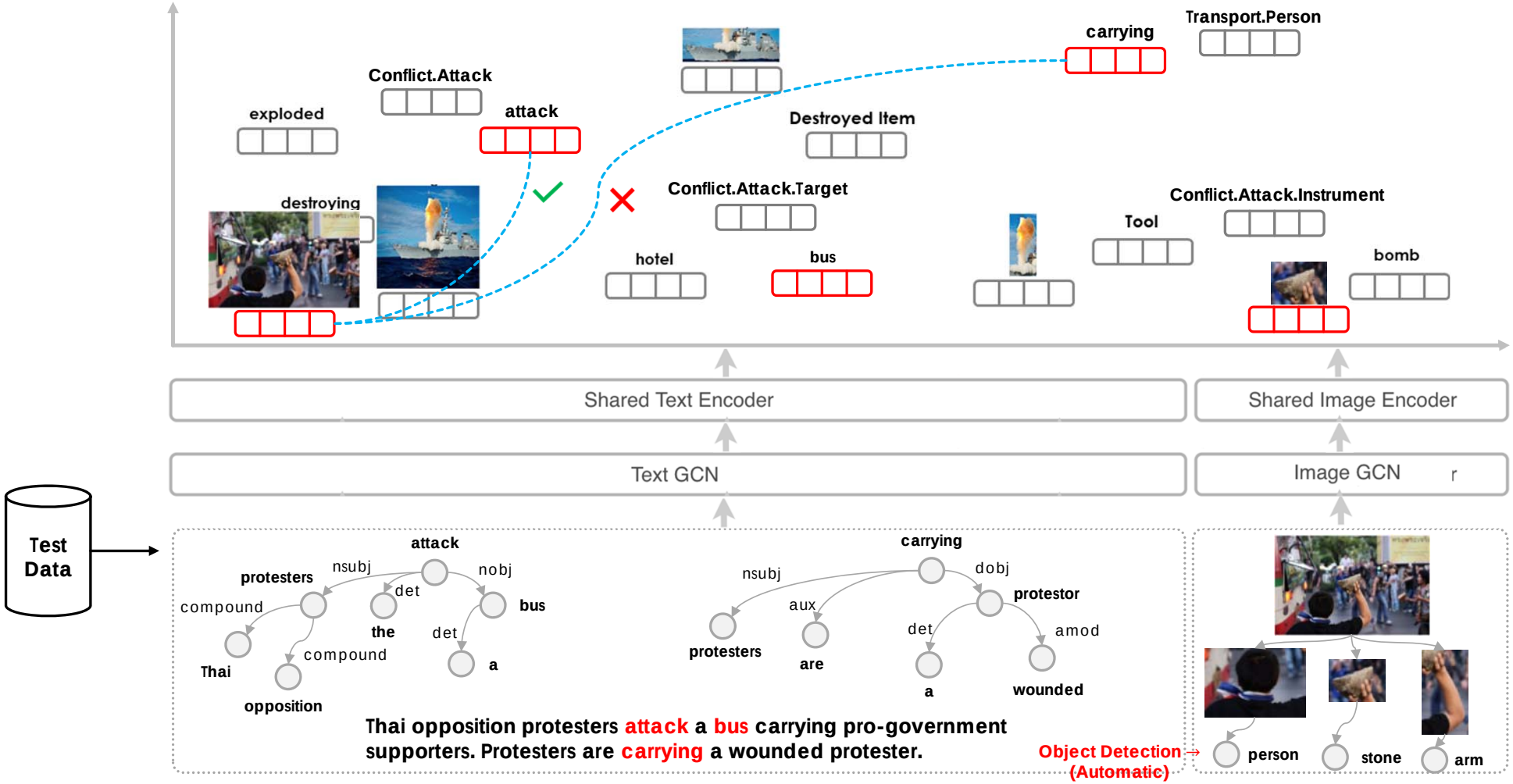
GCN + Encoders for Mapping to Common Space



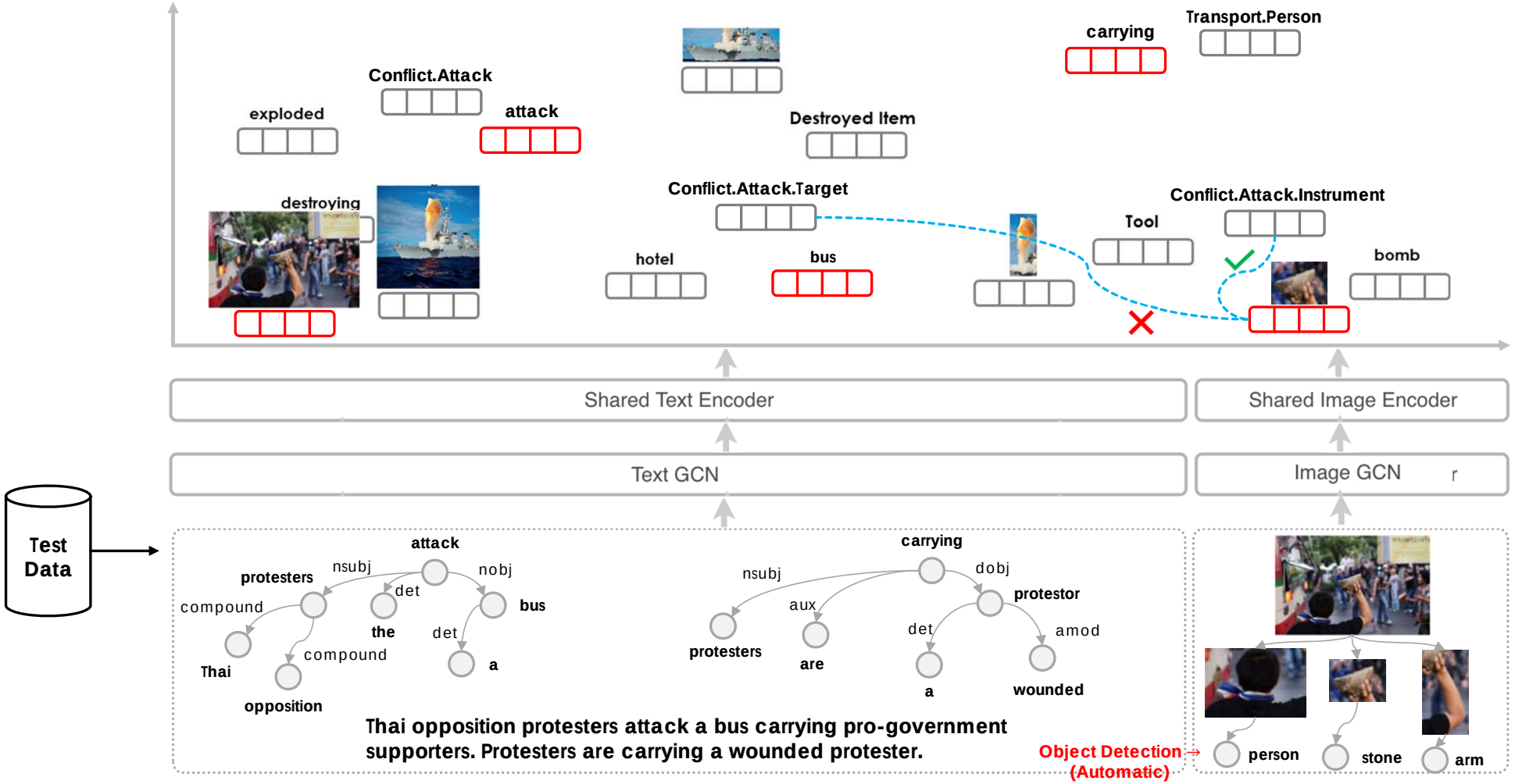
Classify Event Types



Cross-Modal Event Detection



Cross-Modal Argument Role Detection



Multimodal Argument Role: Evaluation

- Experiments
  - Text Event Extraction
    - Input: Text
    - Output: Events
  - Image Event Extraction
    - Input: Image
    - Output: Events
  - Multimedia Event Extraction
    - Input: Text + Image
    - Output: Events

Method	Text Event Extraction		Image Event Extraction		Multimodal Event Extraction	
	Event Type	Argument Role	Event Type	Argument Role	Event	Argument Role
Word Embedding based Semantic Space [1]	2.4%	1.7%	2.0%	2.6%	16%	4.3%
Multimodal Embedding Space (Ours)	45.2%	12.9%	36.0%	9.2%	18.0%	5.6%

Table 1. Accuracy Comparison.

[1] Peters, Matthew, et al. "Deep Contextualized Word Representations." *NAACL*. 2018.

Challenge 3: Multimodal Argument Role

- Sample Image Event Extraction Result

- Baseline:
  - Event: Justice:Arrest-Jail ✗
  - Roles:
    - None ✗
- Our Approach:
  - Event: Conflict.Attack ✓
  - Roles:
    - Instrument = weapon_1 ✓

- Baseline:
  - Event: Justice:Arrest-Jail ✗
  - Roles:
    - Participant = man_1 ✓
- Our Approach:
  - Event: Conflict:Demonstrate ✓
  - Roles:
    - Participant = man_1 ✓

- Baseline:
  - Event: Justice:Arrest-Jail ✓
  - Roles:
    - Agent = man_2 ✗
- Our Approach:
  - Event: Justice:Arrest-Jail ✓
  - Roles:
    - Patient = man_2 ✓

Challenge 3: Multimodal Argument Role

- Challenging Image Event Extraction examples

- Visually similar to other event types
- Ground Truth:
  - Event: Conflict.Demonstrate
  - Roles:
    - Participant = man_4
- Our Approach:
  - Event: Conflict.Attack ✗
  - Roles:
    - Agent = man_4 ✗

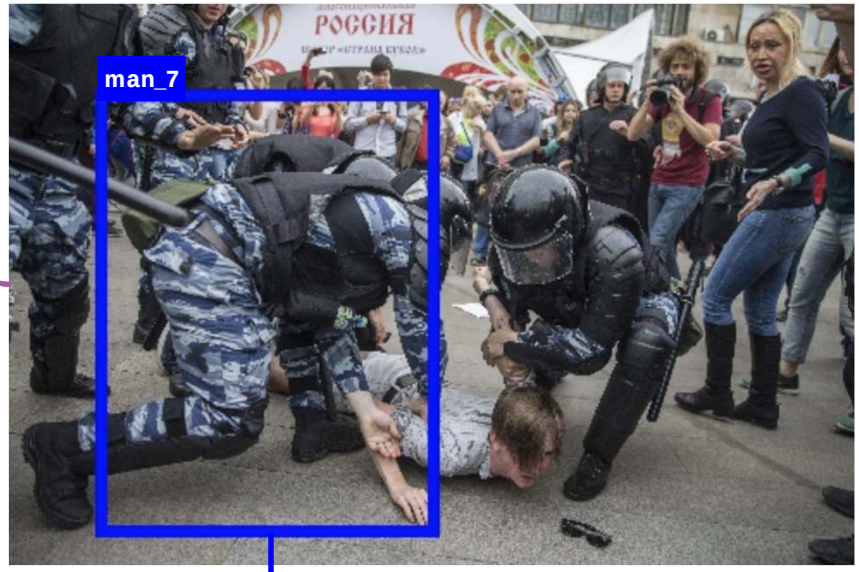
- Irrelevant visual objects to the event
- Ground Truth:
  - Event: Conflict.Demonstrate
  - Roles:
    - None
- Our Approach:
  - Event: Conflict.Demonstrate ✓
  - Roles:
    - Participant = man_5 ✗

- Distinguish semantics, e.g., Attacker (Agent) vs Target (Entity)
- Ground Truth:
  - Event: Conflict.Attack
  - Roles:
    - Agent = man_6
- Our Approach:
  - Event: Conflict.Attack ✓
  - Roles:
    - Entity = man_6 ✗

Challenge 3: Multimodal Argument Role

- Sample Multimedia Event Extraction Result

News Article:
Reporters witnessed hundreds of protesters being detained [Justice.ArrestJail]. More than 3, 000 protested [Conflict.Demonstrate] in the Siberian city of Novosibirsk, with other rallies [Conflict.Demonstrate] in the southern resort Sochi, Krasnoyarsk, Kazan , Tomsk and Vladivostok. Moscow city hall labeled the change in the protest site a provocation and said [Contact.Broadcast] that demonstrations [Conflict.Demonstrate] would be viewed as a threat to public order , leading to the detentions .



- Event matched to image:
  - detain [Justice.ArrestJail] ✓
- Roles from the image:
  - Agent = man_7 ✓

Challenge 3: Multimodal Argument Role

- Challenging Multimedia Event Extraction examples
  - Selecting the correct event instance based on the context

News Article:
Last week , U.S . Secretary of State Rex Tillerson visited [Movement.TransportPerson] Ankara, the first senior administration official to visit [Movement.TransportPerson] Turkey, to try to seal a deal about the battle [Conflict.Attack] for Raqqa and to overcome President Recep Tayyip Erdogan's strong objections to Washington's backing of the Kurdish Democratic Union Party (PYD) militias. Turkish forces have attacked SDF forces in the past around Manbij, west of Raqqa, forcing the United States to deploy [Movement.TransportPerson] dozens of soldiers on the outskirts of the town in a mission to prevent a repeat of clashes, which risk derailing an assault on Raqqa .

- Event matched to image:
  - deploy [Movement.TransportPerson] ✓
- Roles from the image:
  - Vehicle = Land_vehicle_1 ✓
  - Vehicle = Land_vehicle_2 ✓

Challenge 3: Multimodal Argument Role

- Challenging Multimedia Event Extraction examples
  - Coreferential events

News Article:
Police detain [Justice.ArrestJail] a protester in downtown Moscow, Russia, Sunday, March 26, 2017. The demonstrations [Conflict.Demonstrate] took place on the 17th anniversary of the first time Putin was elected [Personnel.Elect] president. Hundreds of peaceful protesters were detained [Justice.ArrestJail] in Moscow, some brutally dragged to the ground just for holding signs criticizing authorities and corruption.

- Event matched to image:
  - detain [Justice.ArrestJail] ✓
  - detained [Justice.ArrestJail] ✓
- Roles from the image:
  - Agent = man_8 ✗

Challenge 3: Multimodal Argument Role

- Challenging Multimedia Event Extraction examples
  - Coreferential events with different trigger words

News Article:
In March , Turkish forces escalated attacks [Conflict.Attack] on the YPG in northern Syria , forcing U.S. to deploy [Movement.TransportPerson] a small number of forces in and around the town of Manbij to the northwest of Raqqa to "deter" Turkish - SDF clashes and ensure the focus remains on Islamic State. Meanwhile, Raqqa is being pummeled by airstrikes [Conflict.Attack] mounted by U.S.-led coalition forces and Syrian warplanes. Local anti-IS activists say the air raids [Conflict.Attack] fail to distinguish between military and non-military targets

- Event matched to image:
  - airstrikes [Conflict.Attack] ✓
  - raids [Conflict.Attack] ✓
- Roles from the image:
  - Target = Airplane_1 ✓
  - Target = Vehicle_1 ✓

Application: Video Description

- Task:
  Given a video data, generate description that depicts visual and relevant knowledge

Whitehead et al. EMNLP'18

Generic:

A man in uniform is talking

Knowledge-aware (by human):

Senior army officer and Zimbabwe Defence Forces' spokesperson, Major General S. B. Moyo, assures the public that President Robert Mugabe and his family are safe and denies that the military is staging a coup.

New Video Description DataSet (Whitehead et al. EMNLP'18)

- New multi-modal dataset introduced
  - News videos with accompanying descriptions and meta-data
  - Descriptions tend to have more newsworthy named entities, compared to existing video captioning datasets

a)

Tags: Independence | Catalonia | Demonstration | Spain

Date: 10/10/2017

Description: Pro-independence supporters gather near the Arc de Triomf in Barcelona to follow the speech of Carles Puigdemont on a big screen.

b)

Tags: Britain | London | Tube | Attack

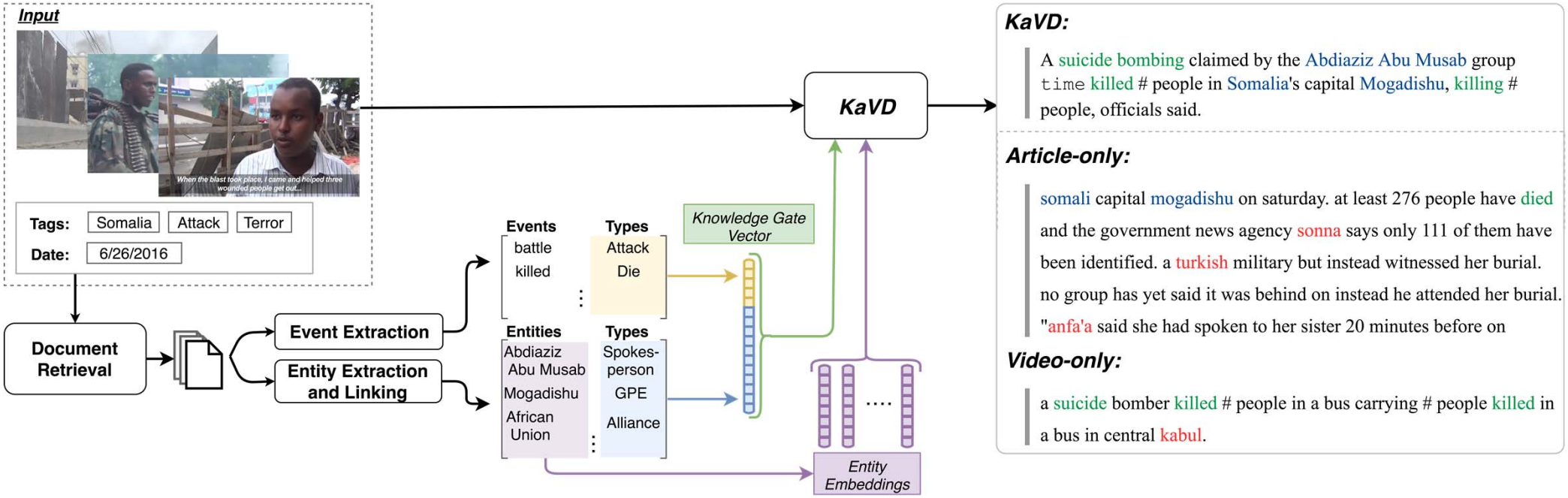
Date: 9/16/2017

Description: There is heightened security on the London Underground Saturday as British police raid a home near London just hours after making their first arrest in the investigation into the bombing of an underground train a day earlier.

Dataset	Domain	#Videos	#Sentences	Vocab Size	Named Entities/Sentence
TACos M-L (Rohrbach et al., 2014)	Cooking	14,105	52,593	2,864	0.1×10^{-4}
MSVD (Chen and Dolan, 2011)	Multi-category	1,970	70,028†	13,010	0.4×10^{-2}
MSR-VTT-10K (Xu et al., 2016)	20 categories	10,000	200,000†	29,316	1.4×10^{-1}
News Video (Ours)	News	2,883	3,302	9,179	2.1

Application: KG Enriched Video Description

- Condition video captioning on knowledge from text and metadata



Whitehead, S., H. Ji, M. Bansal, S.-F. Chang, and C. Voss. "Incorporating Background Knowledge into Video Description Generation." EMNLP, 2018.

Application: KG-Enriched Video Description

- Article-only: not specific to video
- VD: no use of knowledge

Model	METEOR	ROUGE-L	Entity F1	Auto-Entity F1	Event F1	Auto-Event F1
Article-only	8.6	13.2	8.7	8.5	1.9	3.6
VD	9.1	17.9	2.5	1.5	1.0	7.3
VD+Entity Pointer	9.7	18.1	15.3	13.6	5.7	7.0
VD+Knowledge Gate	9.8	18.5	10.2	10.7	6.7	8.3
Entity Pointer+Knowledge Gate	10.1	18.7	**23.7**	**20.9**	2.2	**9.9**
KaVD	**10.2**	**18.9**	22.1	19.7	**9.6**	8.9

Conclusions

- Multimodal Knowledge Graphs
  - Understanding semantic structures in both language and vision
  - Linking across modalities
  - Many challenges and applications
- Open Problems
  - Dynamic knowledge graphs for video
  - Commonsense:
    physics, affordance, behavior, causal/temporal