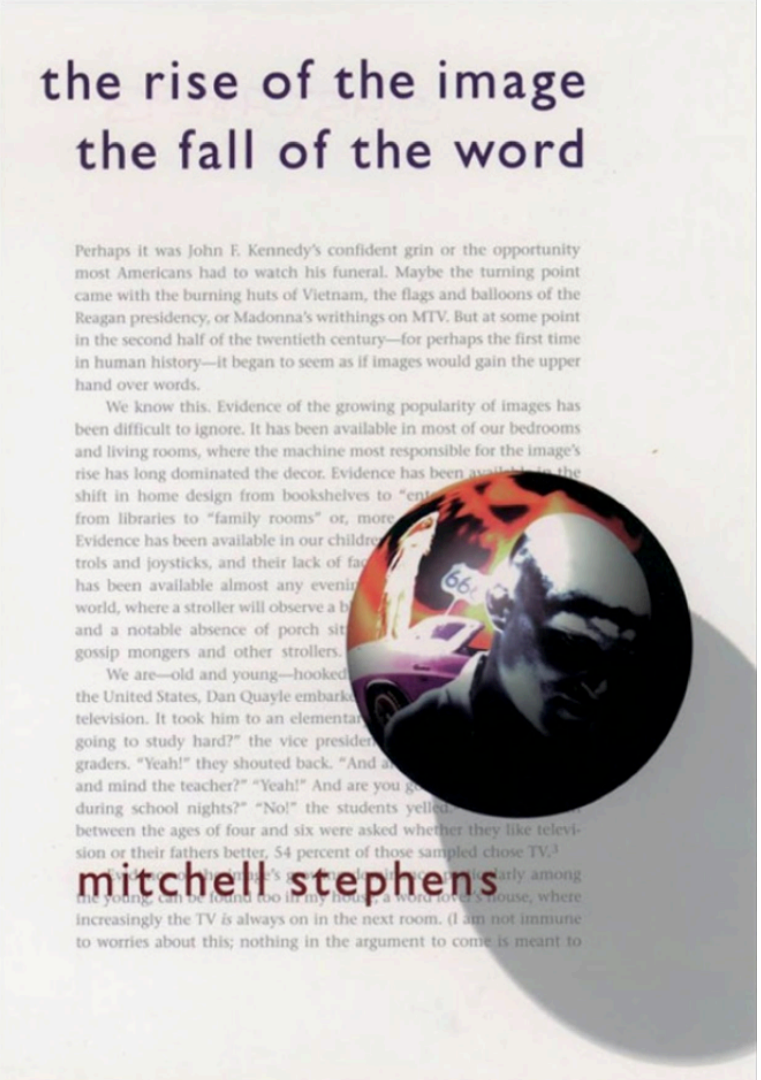
Cross-media Structured Common Space for Multimedia Event Extraction

Manling Li*, Alireza Zareian*, Qi Zeng, Spencer Whitehead, Di Lu,
Heng Ji, Shih-Fu Chang

ACL 2020

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK



Knowledge is Beyond Just Text

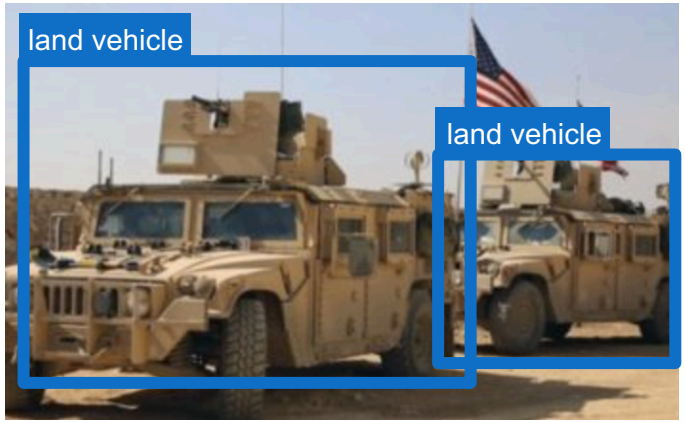
- We produce and consume news content through multimedia, 33% of news images contain event arguments not mentioned in surrounding texts

***TransportPerson_Instrument* = stretcher**

A New Task: Multimedia Event Extraction (M^2E^2)

Input: News Article Text and Image

Last week , U.S . Secretary of State Rex Tillerson visited Ankara, the first senior administration official to visit Turkey, to try to seal a deal about the battle 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** 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.



Output: Events & Argument Roles

Event Type	Movement.Transport	
Event	**Text Trigger**	deploy
	Image	

Arguments	**Agent**	United States
	Destination	outskirts
	Artifact	soldiers
	Vehicle	
	Vehicle	

A New Task: Multimedia Event Extraction (M^2E^2)

Input: News Article Text and Image

Last week , U.S . Secretary of State Rex Tillerson visited Ankara, the first senior administration official to visit Turkey, to try to seal a deal about the battle 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** 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.

Output: Multimedia Events & Argument Roles

Event Type	Movement.Transport			
Event	**Text Trigger**	deploy	**Arguments**	**Agent**	United States
				Destination	outskirts
				Artifact	soldiers
	Image			**Vehicle**	
				Vehicle	

A New Task: Multimedia Event Extraction (M^2E^2)

Input: News Article Text and Image

Last week , U.S . Secretary of State Rex Tillerson visited Ankara, the first senior administration official to visit Turkey, to try to seal a deal about the battle 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** 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.

Output: Multimedia Events & Argument Roles

Event Type	Movement.Transport	
Event	**Text Trigger**	deploy
	Image	

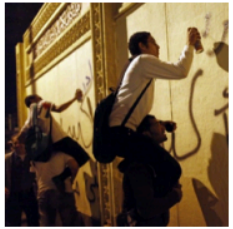
Arguments	**Agent**	United States
	Destination	outskirts
	Artifact	soldiers
	Vehicle	
	Vehicle	

Vision vs. NLP for Event Extraction

- Vision does not study newsworthy, complex events
  - Focusing on daily life and sports (Perera et al., 2012; Chang et al., 2016; Zhang et al., 2007; Ma et al., 2017)
  - Without localizing a complete set of arguments for each event (Gu et al., 2018; Li et al., 2018; Duarte et al., 2018; Sigurdsson et al., 2016; Kato et al., 2018; Wu et al., 2019a)
- Most related: Situation Recognition (Yatskar et al., 2016)
  - Classify an image as one of 500+ FrameNet verbs
  - Identify 192 generic semantic roles via a 1-word description

CLIPPING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

JUMPING			
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR



SPRAYING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

A New Dataset for M^2E^2 Evaluation

- Ontology: intersection between ACE and imSitu
  - **Event Types**: Manually map verbs in imSitu to ACE to obtain the overlapped event types (cover 52% of ACE events in VOA)
  - **Argument Roles**: Based on ACE argument roles, add additional detectable visual roles (marked in red)

Event Type	Argument Roles
Life.Die	Agent, Victim, Instrument, Place
Transaction.TransferMoney	Giver, Recipient, Money, Place
Conflict.Attack	Attacker, Target, Instrument, Place
Conflict.Demonstrate	Entity, Instrument, Police, Place
Contact.Phone-Write	Entity, Instrument, Place
Contact.Meet	Participant, Place
Justice.ArrestJail	Agent, Person, Instrument, Place
Movement.Transport	Agent, Artifact, Vehicle, Destination, Origin

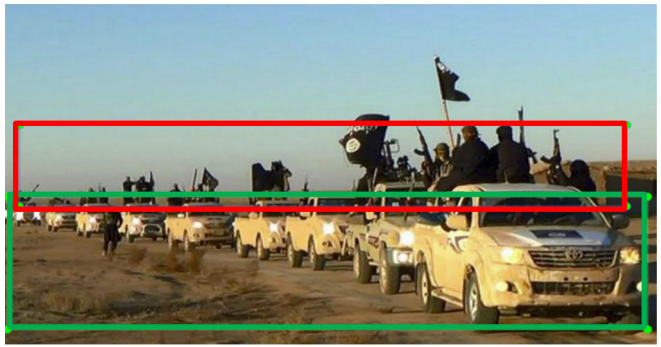
A New Dataset for M^2E^2 Evaluation

- Event type annotation
  - Text: Event Type & Trigger
  - Image: Event Type
- Argument role annotation
  - Text: Argument Role & Entity
  - Image: Argument Role & Bounding Box (Union/Instance Bounding Box)
- Cross-media event coreference resolution
- Two independent annotations + expert annotator adjudication
- Data Source: 245 multimedia news articles from VOA News Website

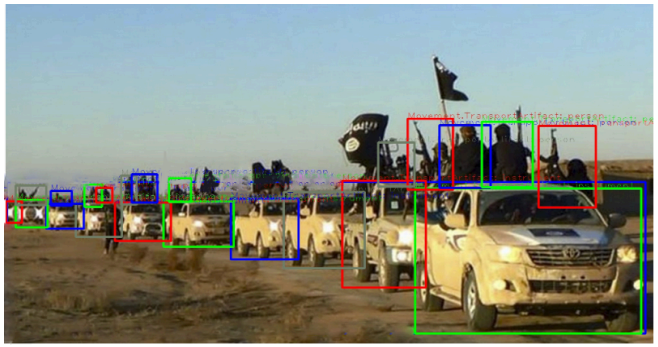
Source Data		Event Mentions		Argument Roles	
# Sentence	# Image	# Textual Mention	# Visual Mention	# Textual Argument	# Visual Argument
6167	1014	1297	391	1965	1429

A New Dataset for M²E² Evaluation

- Bounding box annotation:
  - **Union Bounding Box**: For each role, annotate the smallest bounding box covering all the arguments
  - **Instance Bounding Box**: For each role, annotate multiple bounding boxes, where each bounding box is the smallest region that covers one argument, following VOC2011 Annotation Guidelines[1].



Union Bounding Box



Instance Bounding Box

[1] http://host.robots.ox.ac.uk/pascal/VOC/voc2011/guidelines.html

A New Dataset for M^2E^2 Evaluation

- Each image is annotated by checking the caption as reference context
  - E.g. Captions help to distinguish Movement.Transportation event and Contact.Meet event from Conflict.Demonstration

Migrants are disembarked from the Italian navy ship 'Vega' in the Sicilian harbour of Augusta, southern Italy, May 4, 2015.

Secessionist referendum official Alexander Malyhin holds a document as he speaks to journalists in the eastern Ukrainian city of Luhansk May 12, 2014.

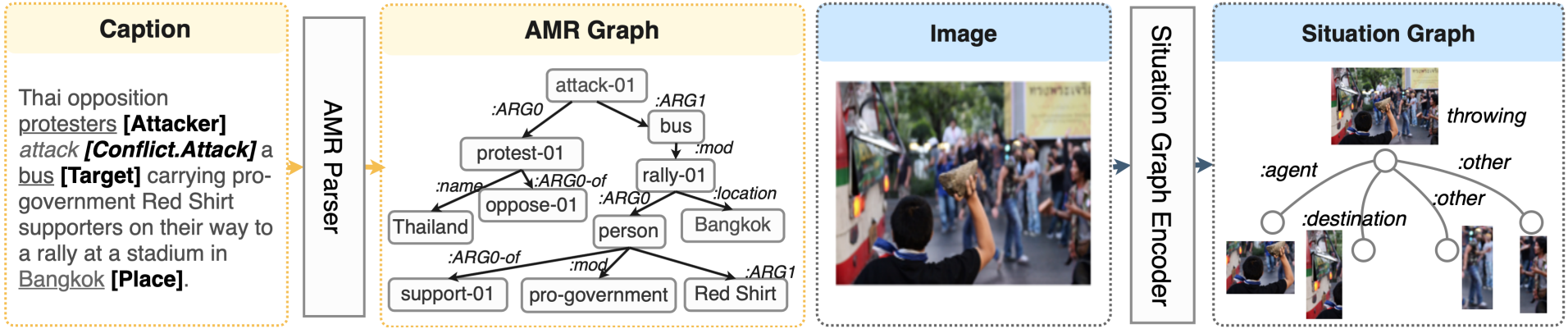
Cross-media Structured Common Space

- Treat Image/Video as a foreign language

Text	Image / Video Frame
Word	Image Region
Entity	Visual Object
Relation	Visual Relation
Entity-Relation Graph	Visual Scene Graph
Event Trigger	Visual Activity
Linguistic Structure	**Situation Graph**

Cross-media Structured Common Space

- Treat Image/Video as a foreign language
  - Represent it with a structure that is similar to AMR graph in text

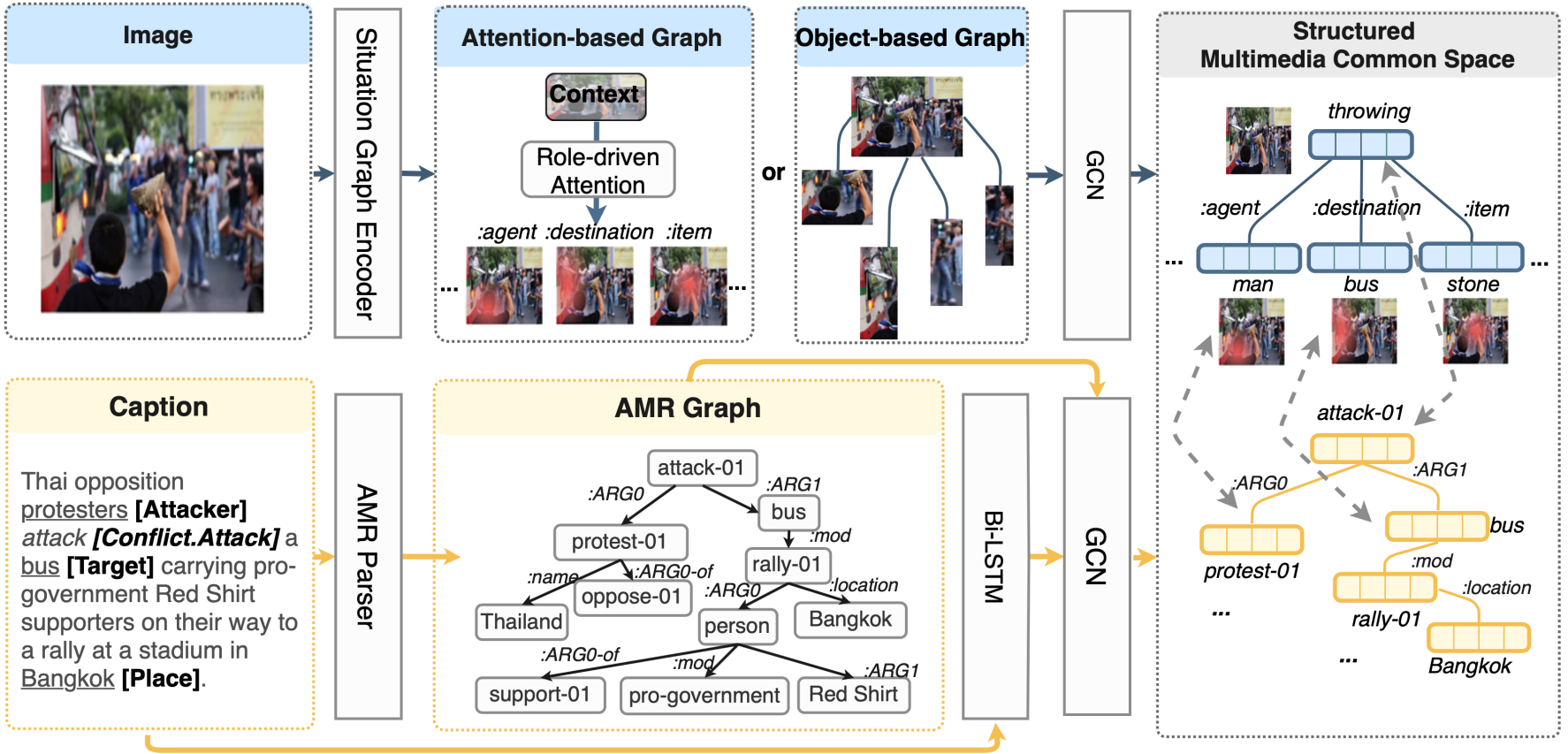


Linguistic Structure,
e.g., Dependency Tree
Abstract Meaning Representation (AMR)

Situation Graph

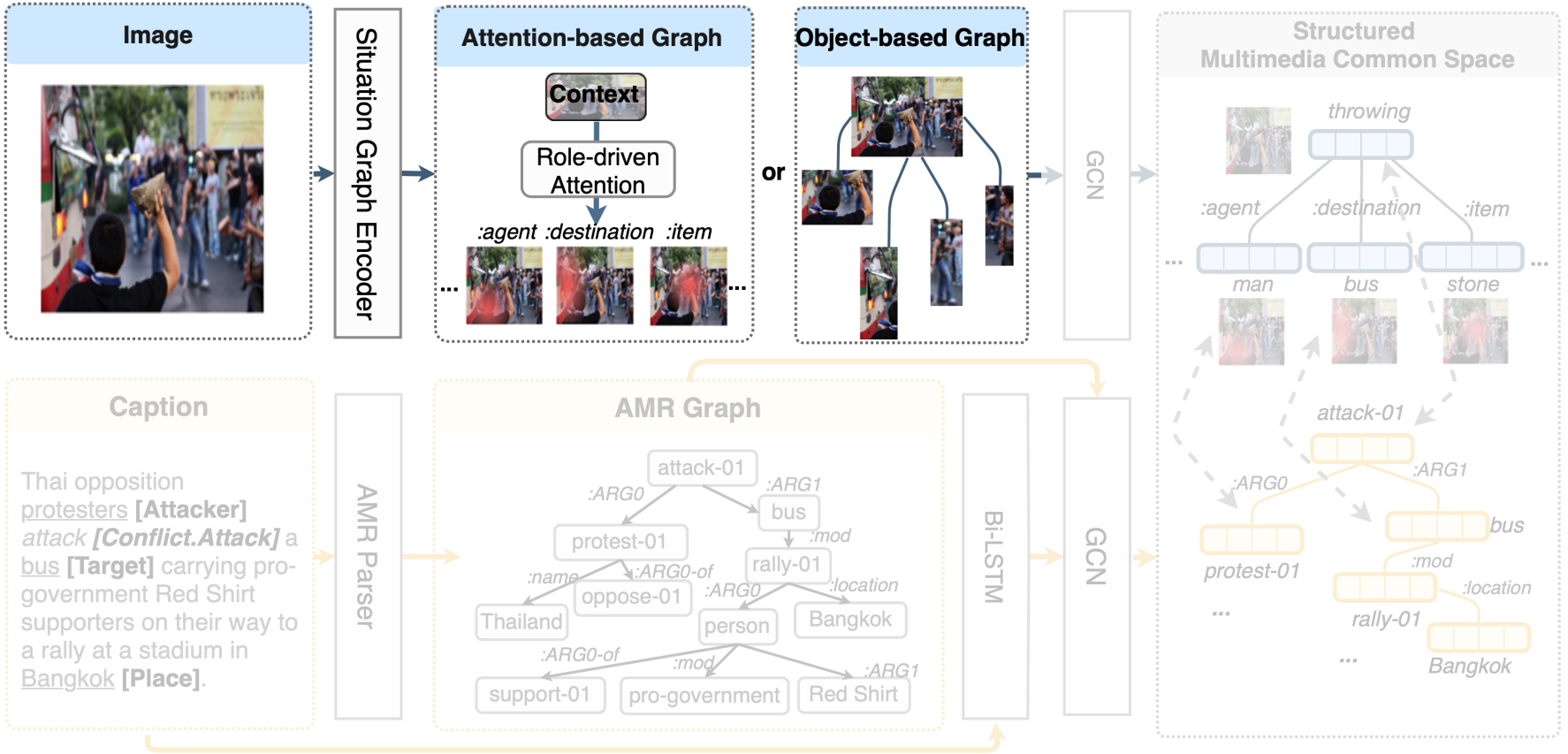
Weakly Aligned Structured Embedding (WASE)

-- Training Phase (Common Space Construction)



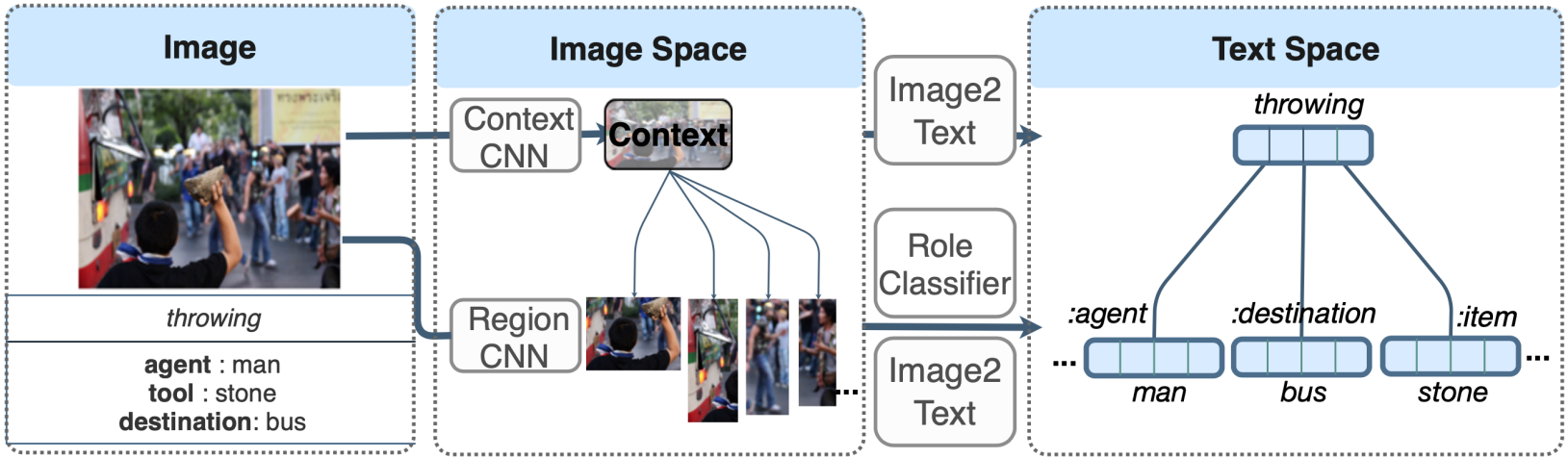
Weakly Aligned Structured Embedding (WASE)

-- Training Phase (Common Space Construction)



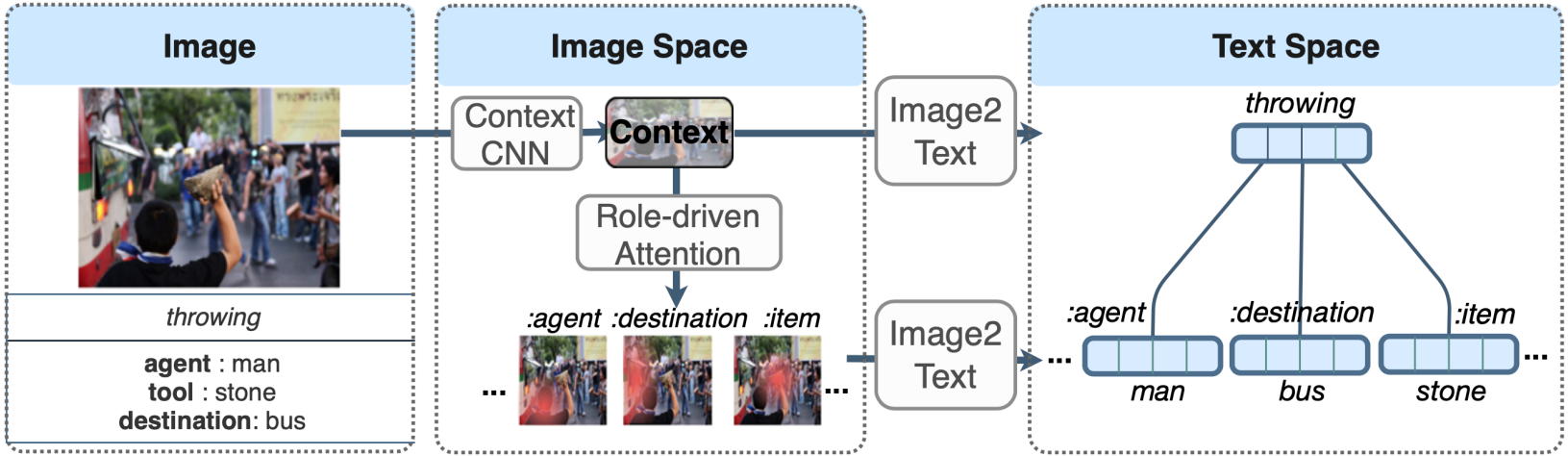
How to generate situation graph?

- Method 1: Object-based Graph Training
  - Learn to project image to verb embedding, and object to noun
  - Learn to classify each object-image pair to a semantic role



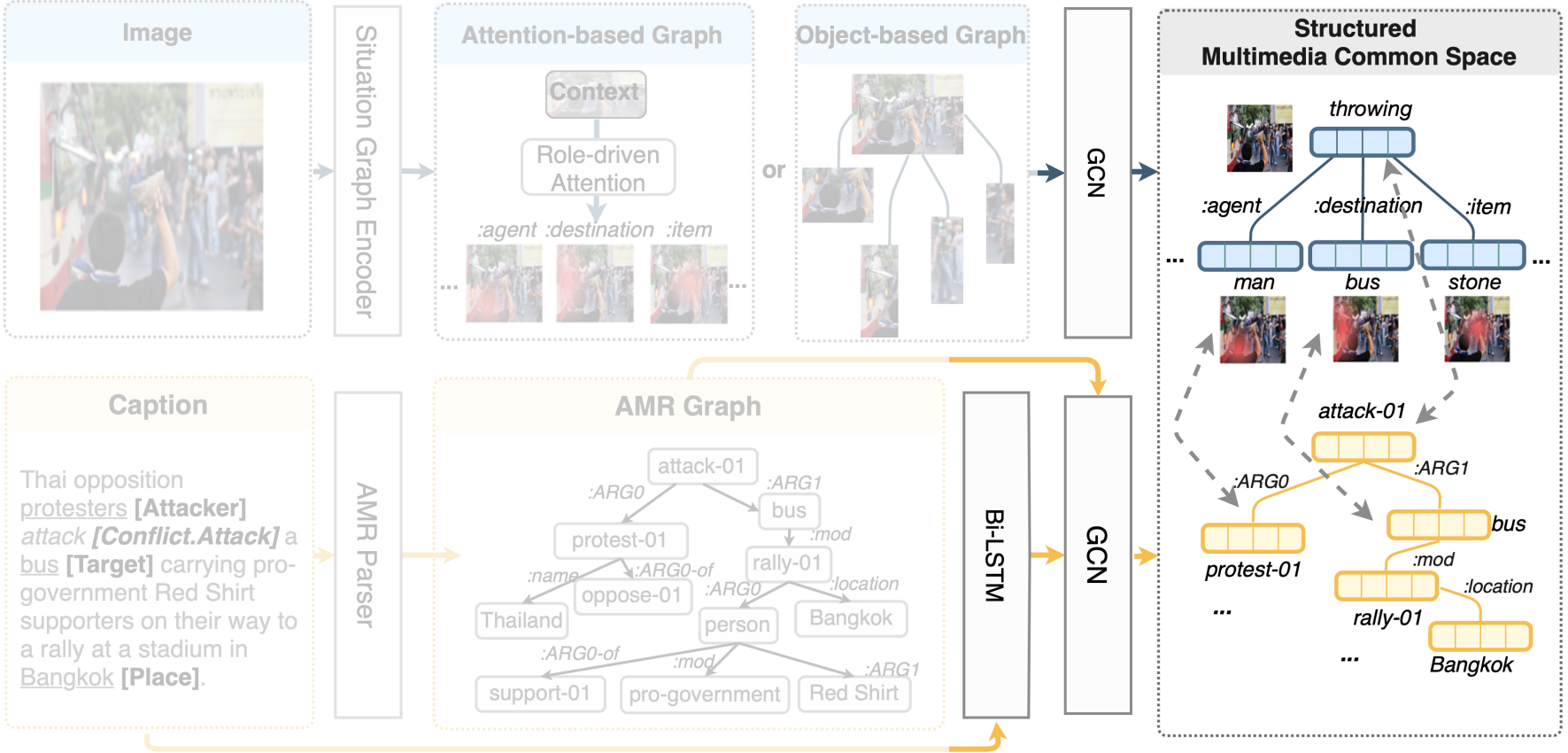
How to generate situation graph?

- Method 2: Role-driven Attention Graph
  - Learn to project image embedding to verb embedding
  - Learn a spatial attention on image for each role
  - Learn to project attended role region to noun embedding



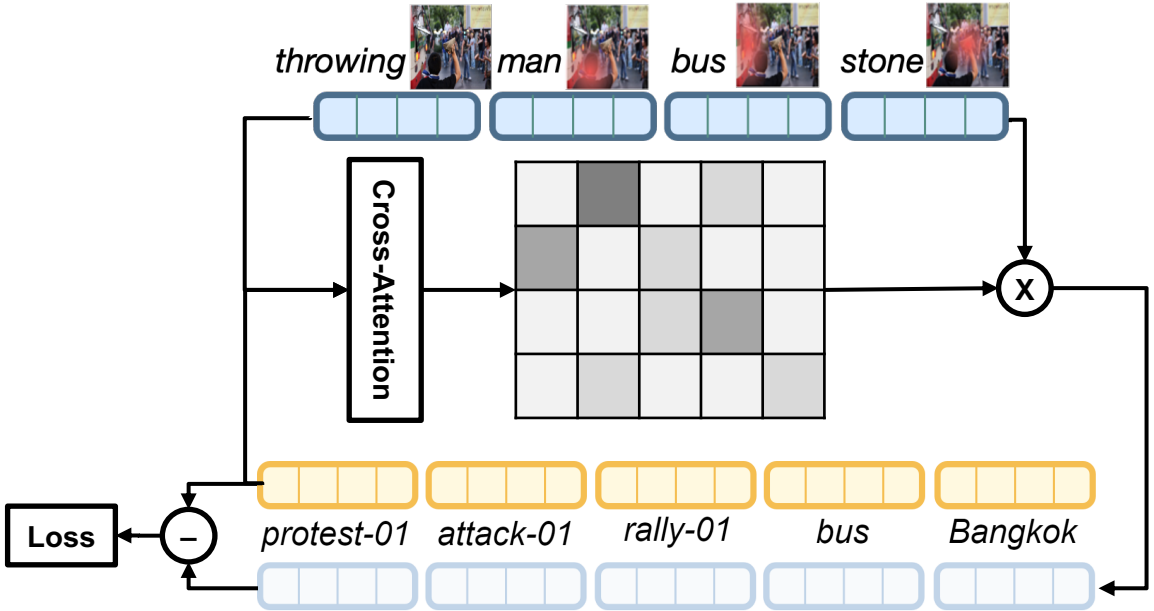
Weakly Aligned Structured Embedding (WASE)

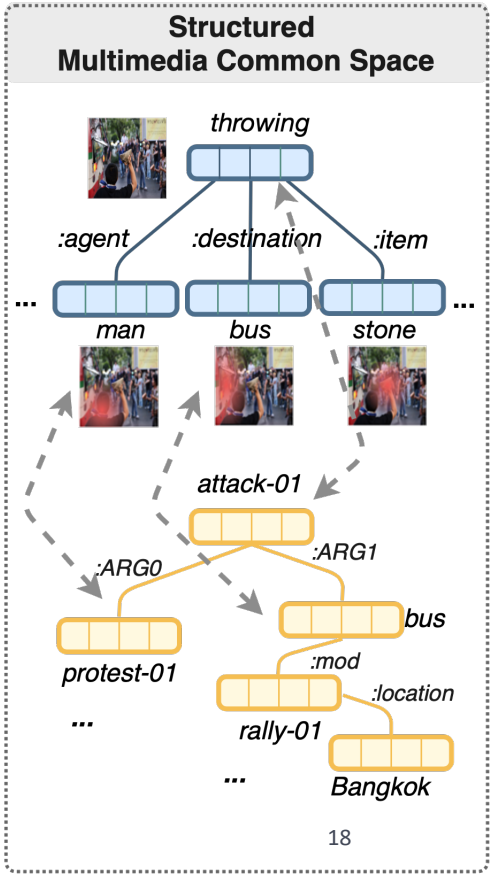
-- Training Phase (Common Space Construction)



How to align the two modalities?

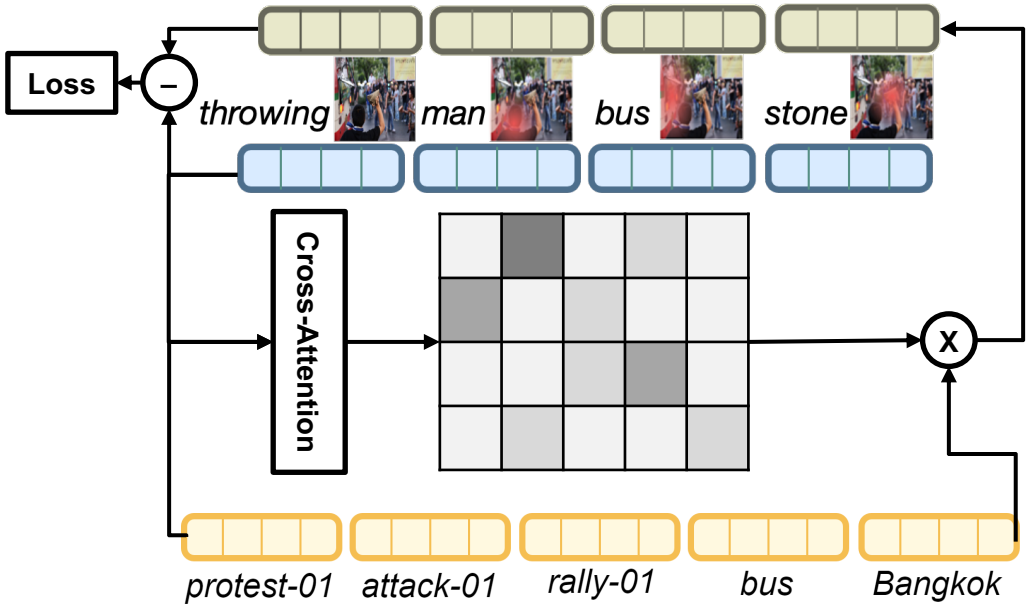
- Prior work aligns image-caption vectors by triplet loss.
- We want to align two graphs, not just single vectors.
- Ontology is shared so the nodes carry similar semantics.

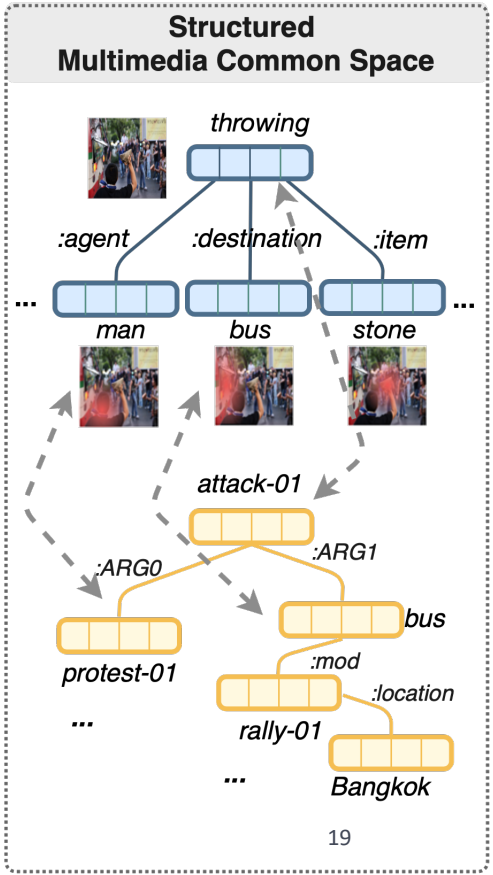




How to align the two modalities?

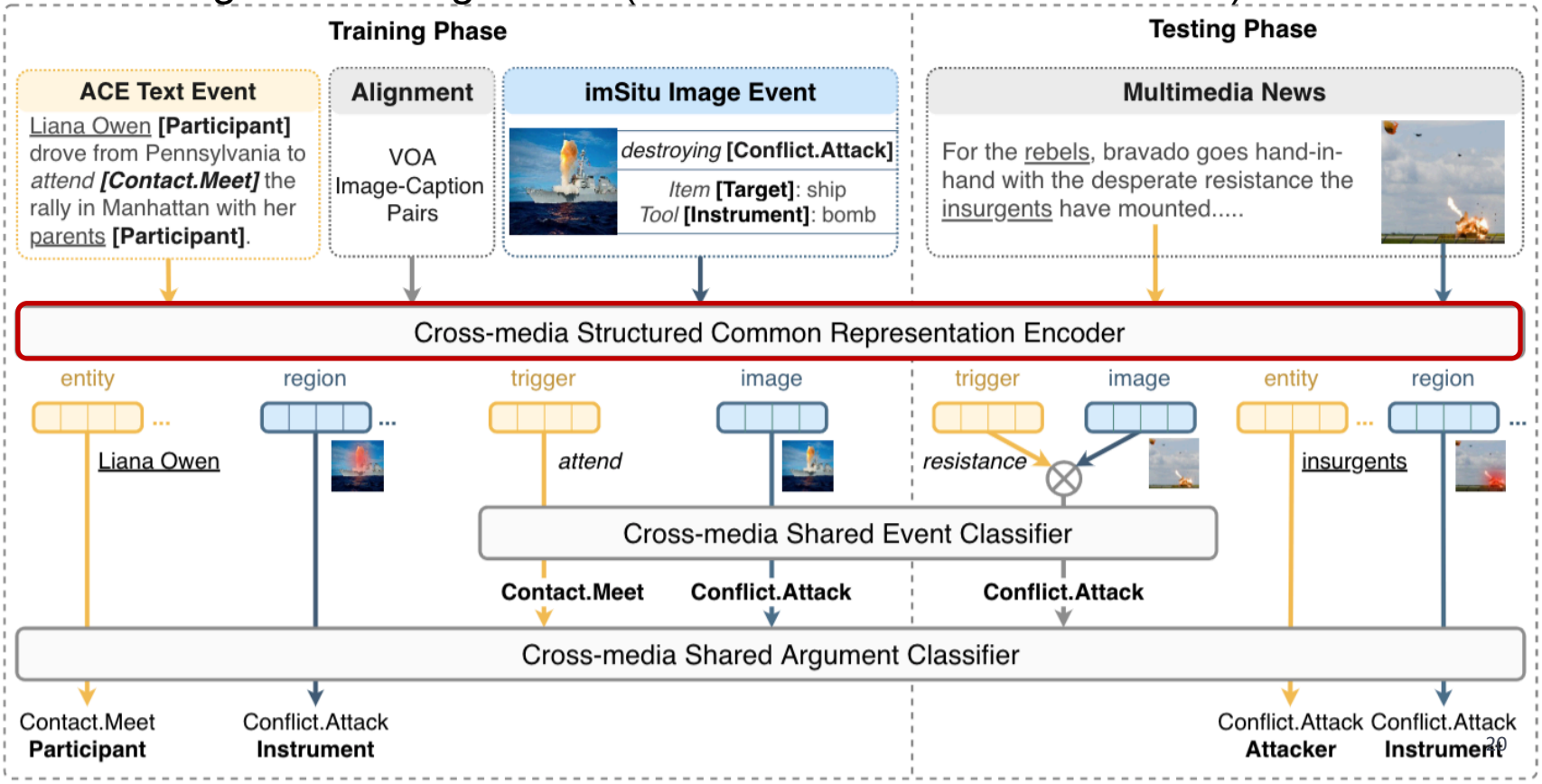
- Prior work aligns image-caption vectors by triplet loss.
- We want to align two graphs, not just single vectors.
- Ontology is shared so the nodes carry similar semantics.





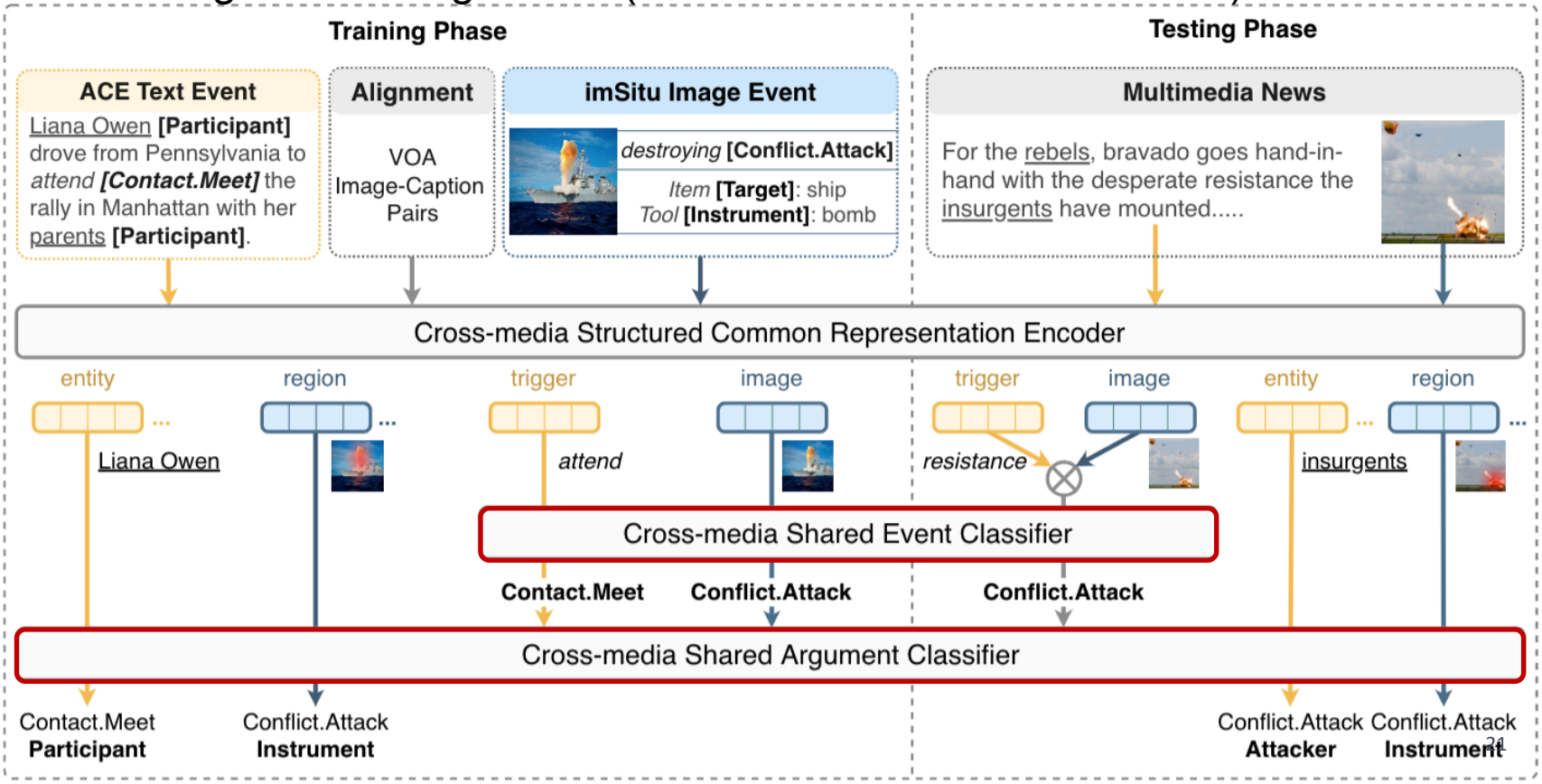
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)



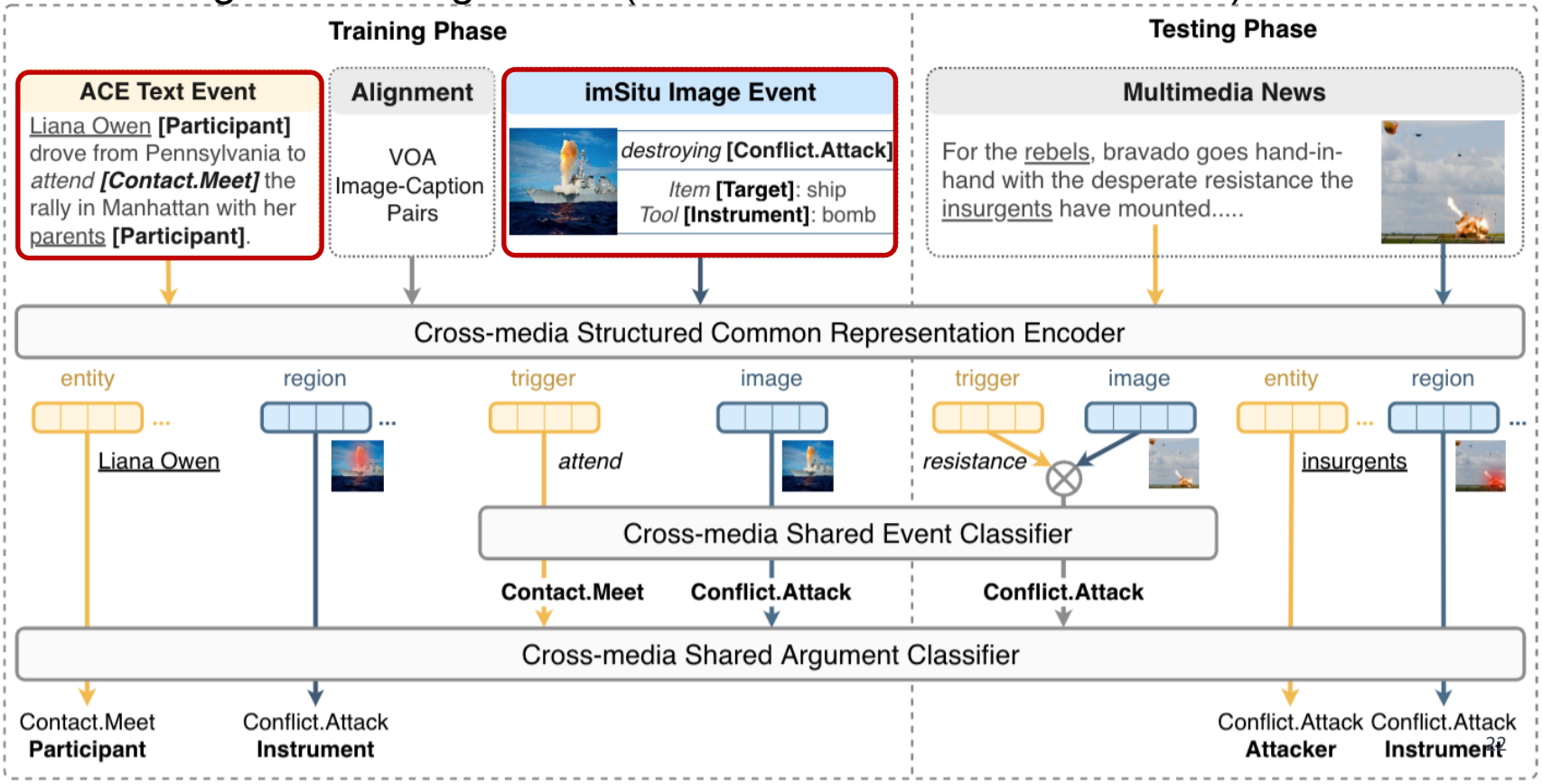
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)



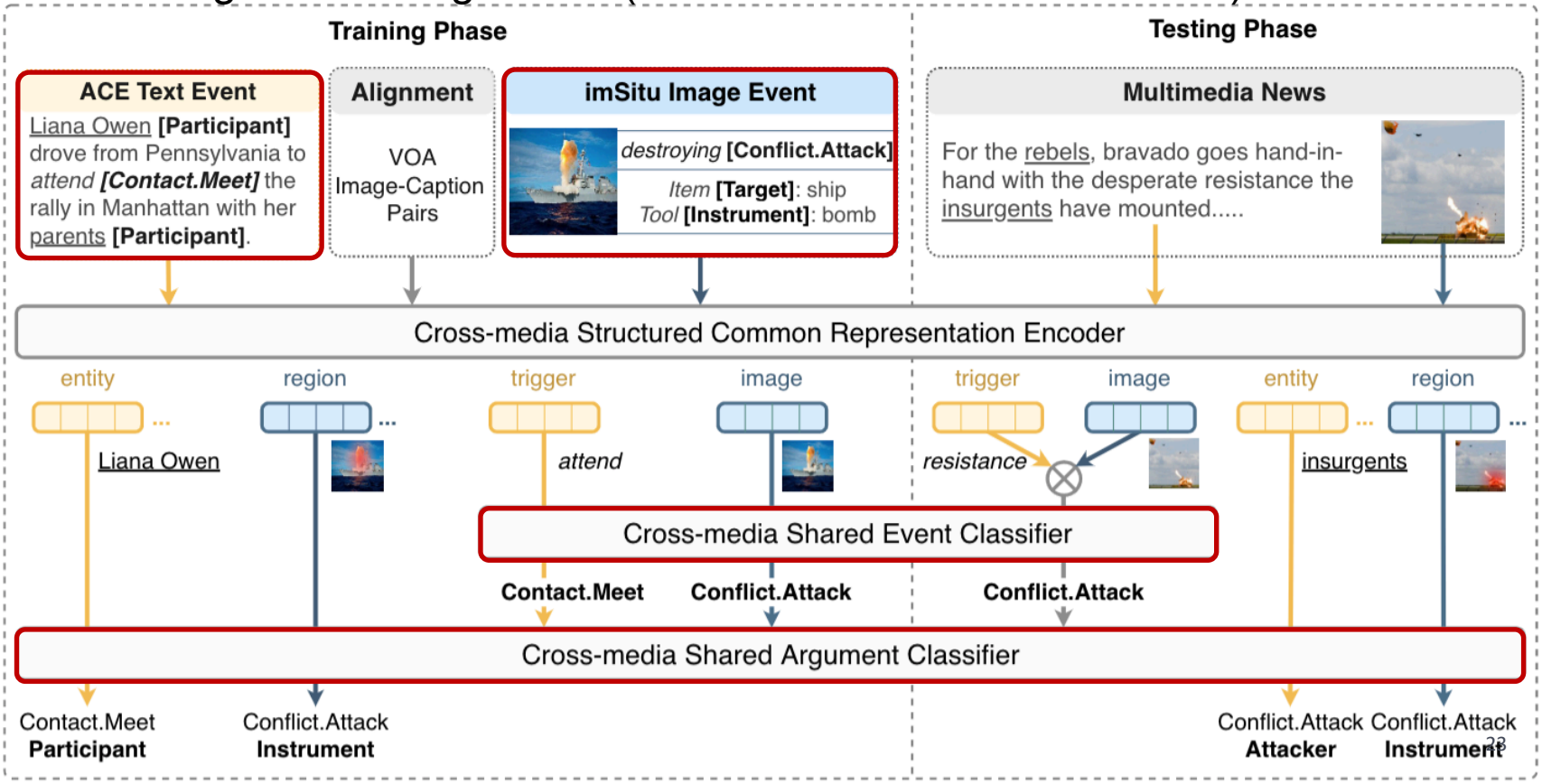
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)



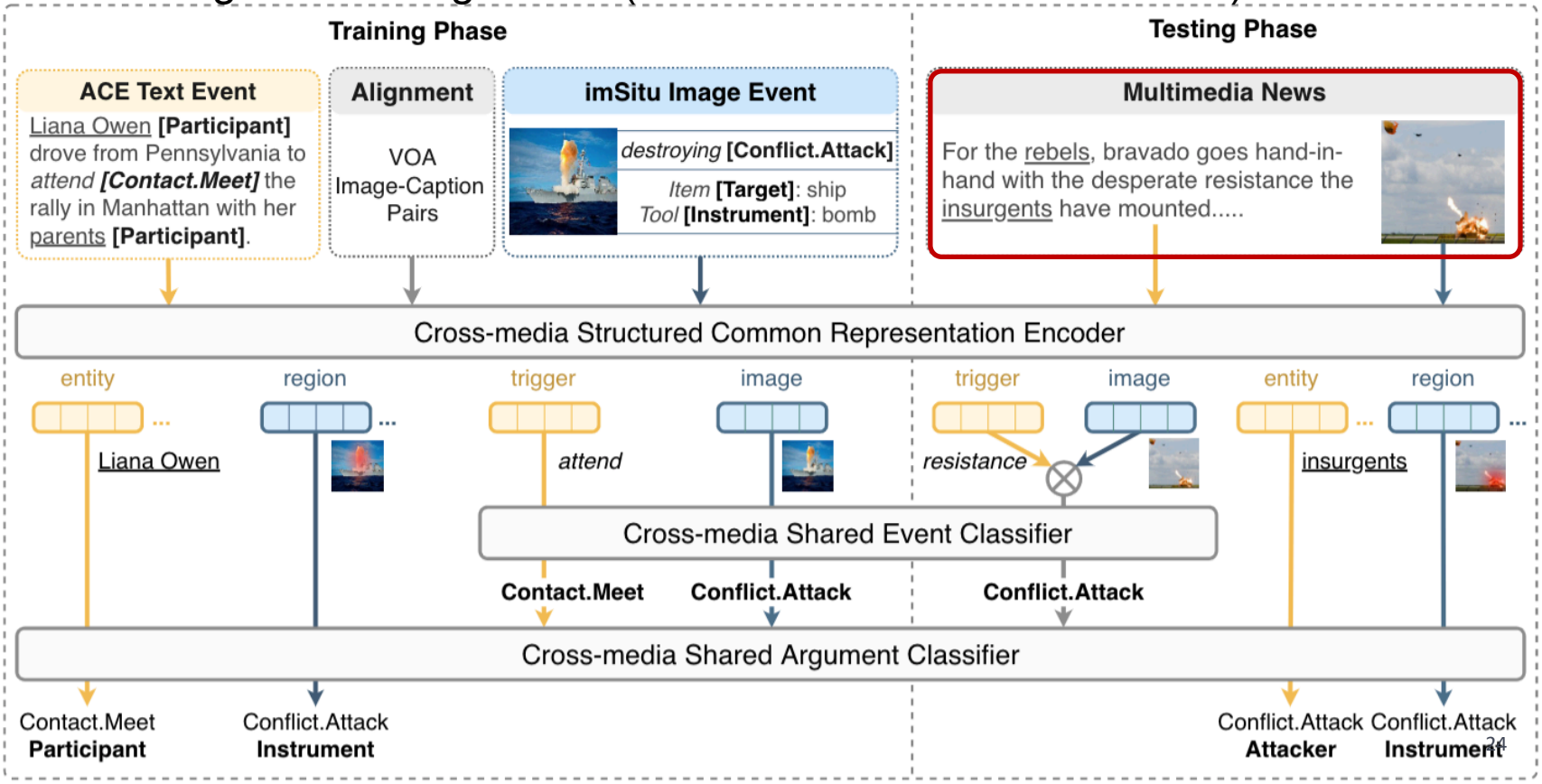
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)



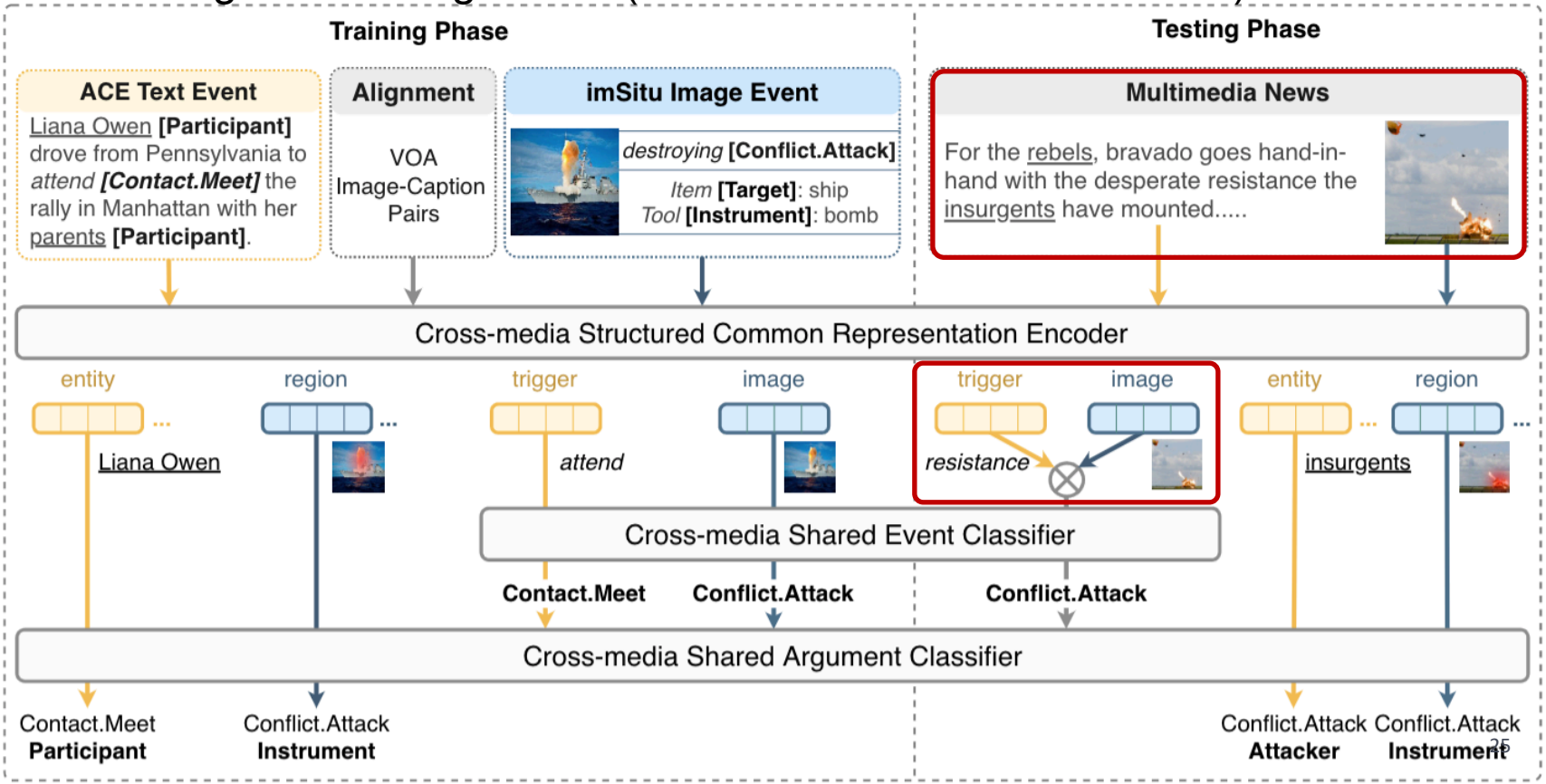
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)



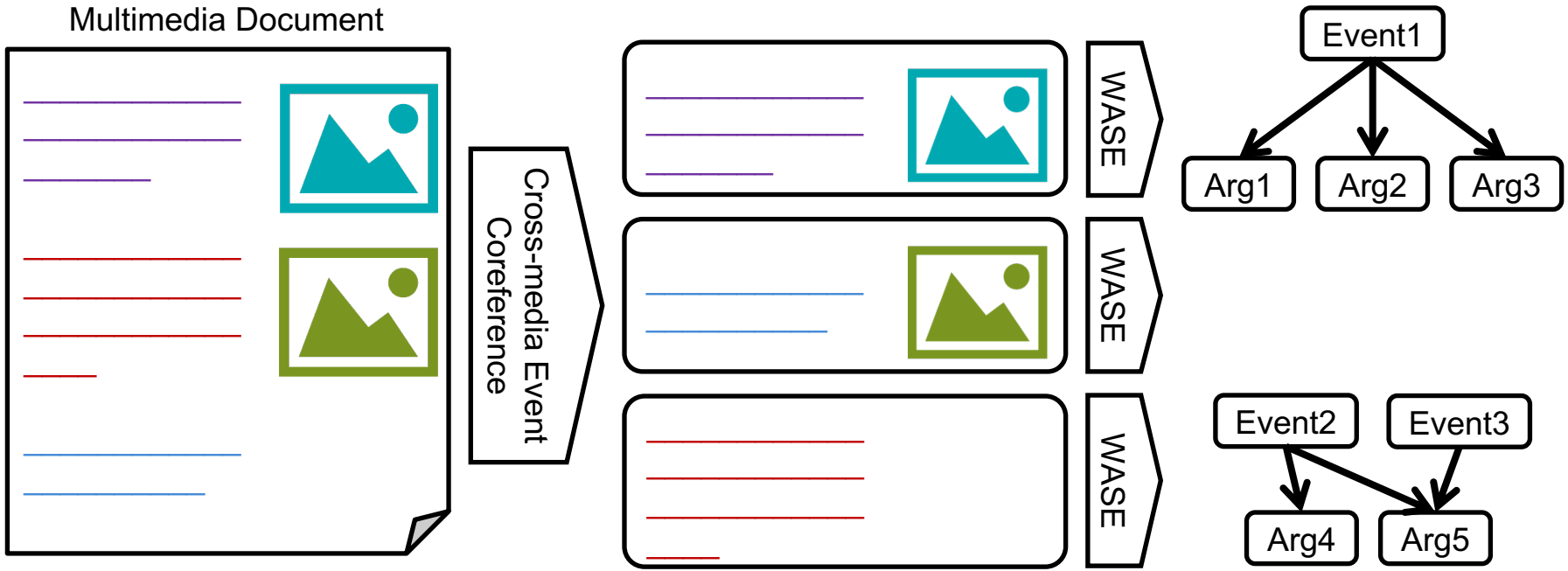
Weakly Aligned Structured Embedding (WASE)

-- Training and Testing Phase (Cross-media shared classifiers)



Weakly Aligned Structured Embedding (WASE)

-- System Diagram



Experiment Results

Training	Model	Text-Only Evaluation						Image-Only Evaluation						Multimedia Evaluation					
		Event Mention			Argument Role			Event Mention			Argument Role			Event Mention			Argument Role		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Text	JMEE	42.5	58.2	48.7	22.9	28.3	25.3	-	-	-	-	-	-	42.1	34.6	38.1	21.1	12.6	15.8
	GAIL	43.4	53.5	47.9	23.6	29.2	26.1	-	-	-	-	-	-	44.0	32.4	37.3	22.7	12.8	16.4
	WASE$^{\mathbb{T}}$	42.3	58.4	48.2	21.4	30.1	24.9	-	-	-	-	-	-	41.2	33.1	36.7	20.1	13.0	15.7
Image	WASE$^{\mathbb{I}}_{\text{att}}$	-	-	-	-	-	-	29.7	61.9	40.1	9.1	10.2	9.6	28.3	23.0	25.4	2.9	6.1	3.8
	WASE$^{\mathbb{I}}_{\text{obj}}$	-	-	-	-	-	-	28.6	59.2	38.7	13.3	9.8	11.2	26.1	22.4	24.1	4.7	5.0	4.9
Multimedia	VSE-C	33.5	47.8	39.4	16.6	24.7	19.8	30.3	48.9	26.4	5.6	6.1	5.7	33.3	48.2	39.3	11.1	14.9	12.8
	Flat$_{\text{att}}$	34.2	63.2	44.4	20.1	27.1	23.1	27.1	57.3	36.7	4.3	8.9	5.8	33.9	59.8	42.2	12.9	17.6	14.9
	Flat$_{\text{obj}}$	38.3	57.9	46.1	21.8	26.6	24.0	26.4	55.8	35.8	9.1	6.5	7.6	34.1	56.4	42.5	16.3	15.9	16.1
	WASE$_{\text{att}}$	37.6	66.8	48.1	27.5	33.2	**30.1**	32.3	63.4	42.8	9.7	11.1	10.3	38.2	67.1	49.1	18.6	21.6	**19.9**
	WASE$_{\text{obj}}$	42.8	61.9	**50.6**	23.5	30.3	26.4	43.1	59.2	**49.9**	14.5	10.1	**11.9**	43.0	62.1	**50.8**	19.5	18.9	19.2

Experiment Results

Training	Model	Text-Only Evaluation						Image-Only Evaluation						Multimedia Evaluation					
		Event Mention			Argument Role			Event Mention			Argument Role			Event Mention			Argument Role		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Text	JMEE	42.5	58.2	48.7	22.9	28.3	25.3	-	-	-	-	-	-	42.1	34.6	38.1	21.1	12.6	15.8
	GAIL	43.4	53.5	47.9	23.6	29.2	26.1	-	-	-	-	-	-	44.0	32.4	37.3	22.7	12.8	16.4
	WASE$^{\mathbb{T}}$	42.3	58.4	48.2	21.4	30.1	24.9	-	-	-	-	-	-	41.2	33.1	36.7	20.1	13.0	15.7
Image	WASE$^{\mathbb{I}}_{\text{att}}$	-	-	-	-	-	-	29.7	61.9	40.1	9.1	10.2	9.6	28.3	23.0	25.4	2.9	6.1	3.8
	WASE$^{\mathbb{I}}_{\text{obj}}$	-	-	-	-	-	-	28.6	59.2	38.7	13.3	9.8	11.2	26.1	22.4	24.1	4.7	5.0	4.9
Multimedia	VSE-C	33.5	47.8	39.4	16.6	24.7	19.8	30.3	48.9	26.4	5.6	6.1	5.7	33.3	48.2	39.3	11.1	14.9	12.8
	Flat$_{\text{att}}$	34.2	63.2	44.4	20.1	27.1	23.1	27.1	57.3	36.7	4.3	8.9	5.8	33.9	59.8	42.2	12.9	17.6	14.9
	Flat$_{\text{obj}}$	38.3	57.9	46.1	21.8	26.6	24.0	26.4	55.8	35.8	9.1	6.5	7.6	34.1	56.4	42.5	16.3	15.9	16.1
	WASE$_{\text{att}}$	37.6	66.8	48.1	27.5	33.2	**30.1**	32.3	63.4	42.8	9.7	11.1	10.3	38.2	67.1	49.1	18.6	21.6	**19.9**
	WASE$_{\text{obj}}$	42.8	61.9	**50.6**	23.5	30.3	26.4	43.1	59.2	**49.9**	14.5	10.1	**11.9**	43.0	62.1	**50.8**	19.5	18.9	19.2

Experiment Results

Training	Model	Text-Only Evaluation						Image-Only Evaluation						Multimedia Evaluation					
		Event Mention			Argument Role			Event Mention			Argument Role			Event Mention			Argument Role		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Text	JMEE	42.5	58.2	48.7	22.9	28.3	25.3	-	-	-	-	-	-	42.1	34.6	38.1	21.1	12.6	15.8
	GAIL	43.4	53.5	47.9	23.6	29.2	26.1	-	-	-	-	-	-	44.0	32.4	37.3	22.7	12.8	16.4
	WASE$^{\mathbb{T}}$	42.3	58.4	48.2	21.4	30.1	24.9	-	-	-	-	-	-	41.2	33.1	36.7	20.1	13.0	15.7
Image	WASE$^{\mathbb{I}}_{\text{att}}$	-	-	-	-	-	-	29.7	61.9	40.1	9.1	10.2	9.6	28.3	23.0	25.4	2.9	6.1	3.8
	WASE$^{\mathbb{I}}_{\text{obj}}$	-	-	-	-	-	-	28.6	59.2	38.7	13.3	9.8	11.2	26.1	22.4	24.1	4.7	5.0	4.9
Multimedia	VSE-C	33.5	47.8	39.4	16.6	24.7	19.8	30.3	48.9	26.4	5.6	6.1	5.7	33.3	48.2	39.3	11.1	14.9	12.8
	Flat$_{\text{att}}$	34.2	63.2	44.4	20.1	27.1	23.1	27.1	57.3	36.7	4.3	8.9	5.8	33.9	59.8	42.2	12.9	17.6	14.9
	Flat$_{\text{obj}}$	38.3	57.9	46.1	21.8	26.6	24.0	26.4	55.8	35.8	9.1	6.5	7.6	34.1	56.4	42.5	16.3	15.9	16.1
	WASE$_{\text{att}}$	37.6	66.8	48.1	27.5	33.2	**30.1**	32.3	63.4	42.8	9.7	11.1	10.3	38.2	67.1	49.1	18.6	21.6	**19.9**
	WASE$_{\text{obj}}$	42.8	61.9	**50.6**	23.5	30.3	26.4	43.1	59.2	**49.9**	14.5	10.1	**11.9**	43.0	62.1	**50.8**	19.5	18.9	19.2

Cross-Media Coreference Accuracy

Model	P (%)	R (%)	F_1 (%)
rule_based	10.1	100	18.2
VSE	31.2	74.5	44.0
Flat_{att}	33.1	73.5	45.6
Flat_{obj}	34.3	76.4	47.3
WASE_{att}	39.5	73.5	51.5
WASE_{obj}	40.1	75.4	52.4

Compare to Single Data Modality Extraction

- Surrounding sentence helps visual event extraction.

People celebrate Supreme Court ruling on Same Sex Marriage in front of the Supreme Court in Washington.

- Image helps textual event extraction.

Iraqi security forces *search* ***[Justice.Arrest]*** a civilian in the city of Mosul.

Why Does Vision Help NLP?

- Various triggers and context can be coherent in visual space.
- Cross-media Common space pushes scattered sentences towards the visual cluster.

Berlin police tweeted that six people were arrested after a joint operation with the Berlin's prosecutor's office.

The man in Kosovo is an ethnic Albanian arrested south of the capital, Pristina.

He was asleep in a suburban Seattle house last week morning when immigration agents showed up to arrest his father.

But shortly after the round table began, Marko Djuric, head of the Serbian government office on Kosovo, was detained by police.

Compare to Cross-media Flat Representation

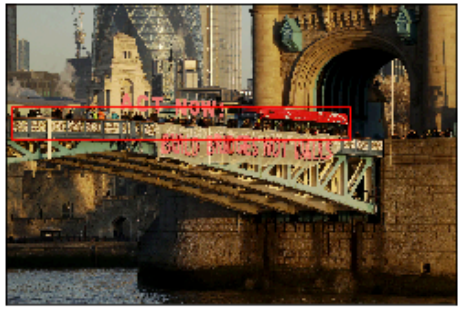
Model	Event Type	Argument Role	
Flat	Justice.ArrestJail	Agent =	man
Ours	Justice.ArrestJail	Entity =	man

Model	Event Type	Argument Role	
Flat	Movement.Transport	Artifact =	none
Ours	Movement.Transport	Artifact =	man

Remaining Challenges: wrong localization

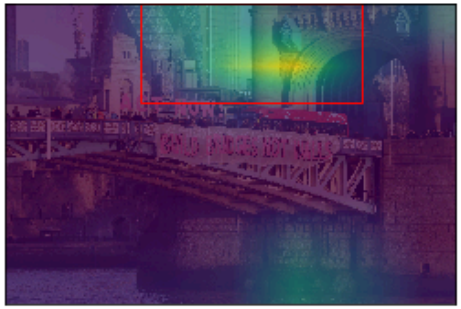
Predicted event: CONFLICT||DEMONSTRATE, ground truth: CONFLICT||DEMONSTRATE
Predicted verb: parading

ENTITY: people



PLACE: open air

ENTITY: demonstrator



Predicted event: CONFLICT||DEMONSTRATE, ground truth: CONTACT||MEET
Predicted verb: parading

ENTITY: people

PLACE: street

ENTITY: troops

Remaining Challenges: too many instances

```
Predicted event: CONFLICT||DEMONSTRATE, ground truth: CONFLICT||DEMONSTRATE
Predicted verb: parading
```

ENTITY: people

PLACE: street

ENTITY: dissent

Conclusion

- A new task, *MultiMedia Event Extraction*, with a evaluation **benchmark**
- A **weakly supervised** training framework, which utilizes existing single-modal annotated corpora, and enables joint inference without cross-modal annotation
- A **structured multimedia common space** to leverage structured representations and graph-based neural networks
- Future Work
  - Extend event extraction to videos
  - Enrich event types
    - Extend to other text event ontologies
    - Discover new event types not in existing text ontologies using zero-shot learning
  - Apply multimedia common semantic space to improve cross-media event and entity coreference resolution, cross-media event inference, event prediction, etc.

Reference

- Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5534–5542.
- Xiaojun Chang, Zhigang Ma, Yi Yang, Zhiqiang Zeng, and Alexander G Hauptmann. 2016. Bilevel semantic representation analysis for multimedia event detection. IEEE transactions on cybernetics, 47(5):1180–1197
- Tongtao Zhang, Spencer Whitehead, Hanwang Zhang, Hongzhi Li, Joseph Ellis, Lifu Huang, Wei Liu, Heng Ji, and Shih-Fu Chang. 2017. Improving event extraction via multimodal integration. In Proceedings of the 25th ACM international conference on Multimedia, pages 270–278. ACM.
- Zhigang Ma, Xiaojun Chang, Zhongwen Xu, Nicu Sebe, and Alexander G Hauptmann. 2017. Joint attributes and event analysis for multimedia event detection. IEEE transactions on neural networks and learning systems, 29(7):2921–2930.
- AG Amitha Perera, Sangmin Oh, P Megha, Tianyang Ma, Anthony Hoogs, Arash Vahdat, Kevin Cannons, Greg Mori, Scott Mccloskey, Ben Miller, et al. 2012. Trecvid 2012 genie: Multimedia event detection and recounting. In In TRECVID Workshop. Citeseer
- Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions.
- Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. 2018. Recurrent tubelet proposal and recognition networks for action detection. In Proceedings of the European conference on computer vision (ECCV), pages 303–318.
- Kevin Duarte, Yogesh Rawat, and Mubarak Shah. 2018. Videocapsulenet: A simplified network for action detection. In Advances in Neural Information Processing Systems, pages 7610–7619.
- Gunnar A Sigurdsson, Gul Varol, Xiaolong Wang, Ali ¨ Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In European Conference on Computer Vision, pages 510–526. Springer
- Keizo Kato, Yin Li, and Abhinav Gupta. 2018. Compositional learning for human object interaction. In Proceedings of the European Conference on Computer Vision (ECCV), pages 234–251.
- Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019a. Long-term feature banks for detailed video understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 284–293.

THANK YOU


ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN


COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK