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
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.
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

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Movement.Transport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>Text Trigger: deploy</td>
</tr>
<tr>
<td></td>
<td>Image</td>
</tr>
</tbody>
</table>

Arguments

<table>
<thead>
<tr>
<th>Agent</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Destination</td>
<td>outskirts</td>
</tr>
<tr>
<td>Artifact</td>
<td>soldiers</td>
</tr>
<tr>
<td>Vehicle</td>
<td>Vehicle</td>
</tr>
</tbody>
</table>
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**

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Movement.Transport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>Text Trigger</td>
</tr>
<tr>
<td></td>
<td>deploy</td>
</tr>
<tr>
<td>Image</td>
<td></td>
</tr>
</tbody>
</table>

**Arguments**

- **Agent**
  - United States
- **Destination**
  - outskirts
- **Artifact**
  - soldiers
- **Vehicle**
  - land vehicle
  - land 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

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Movement.Transport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event</td>
<td>Text Trigger deploy</td>
</tr>
<tr>
<td>Image</td>
<td>United States soldiers</td>
</tr>
</tbody>
</table>

Arguments

- Agent: United States
- Destination: outskirts
- Artifact: soldiers
- Vehicle: land 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
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)

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Argument Roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life.Die</td>
<td>Agent, Victim, Instrument, Place</td>
</tr>
<tr>
<td>Transaction.TransferMoney</td>
<td>Giver, Recipient, Money, Place</td>
</tr>
<tr>
<td>Conflict.Attack</td>
<td>Attacker, Target, Instrument, Place</td>
</tr>
<tr>
<td>Conflict.Demonstrate</td>
<td>Entity, <strong>Instrument</strong>, Police, Place</td>
</tr>
<tr>
<td>Contact.Phone-Write</td>
<td>Entity, <strong>Instrument</strong>, Place</td>
</tr>
<tr>
<td>Contact.Meet</td>
<td>Participant, Place</td>
</tr>
<tr>
<td>Justice.ArrestJail</td>
<td>Agent, Person, <strong>Instrument</strong>, Place</td>
</tr>
<tr>
<td>Movement.Transport</td>
<td>Agent, Artifact, Vehicle, Destination, Origin</td>
</tr>
</tbody>
</table>
A New Dataset for M$^2$E$^2$ 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

<table>
<thead>
<tr>
<th>Source Data</th>
<th>Event Mentions</th>
<th>Argument Roles</th>
</tr>
</thead>
<tbody>
<tr>
<td># Sentence</td>
<td># Image</td>
<td># Textual Mention</td>
</tr>
<tr>
<td>6167</td>
<td>1014</td>
<td>1297</td>
</tr>
</tbody>
</table>
A New Dataset for M$^2$E$^2$ 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$^1$.

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

<table>
<thead>
<tr>
<th>Text</th>
<th>Image / Video Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Image Region</td>
</tr>
<tr>
<td>Entity</td>
<td>Visual Object</td>
</tr>
<tr>
<td>Relation</td>
<td>Visual Relation</td>
</tr>
<tr>
<td>Entity-Relation Graph</td>
<td>Visual Scene Graph</td>
</tr>
<tr>
<td>Event Trigger</td>
<td>Visual Activity</td>
</tr>
<tr>
<td><strong>Linguistic Structure</strong></td>
<td><strong>Situation Graph</strong></td>
</tr>
</tbody>
</table>
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)
Weakly Aligned Structured Embedding (WASE)

-- Training Phase (Common Space Construction)

Caption
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

![Diagram showing the process of generating situation graphs from images and text]
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

```
<table>
<thead>
<tr>
<th>Image</th>
<th>Image Space</th>
<th>Text Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>throwing</td>
<td>Context CNN</td>
<td>throwing</td>
</tr>
<tr>
<td>agent: man</td>
<td>Role-driven Attention</td>
<td>:agent</td>
</tr>
<tr>
<td>tool: stone</td>
<td></td>
<td>:destination</td>
</tr>
<tr>
<td>destination</td>
<td></td>
<td>:item</td>
</tr>
<tr>
<td>bus</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

16
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)

ACE Text Event
Liana Owen [Participant] drove from Pennsylvania to attend [Contact.Meet] the rally in Manhattan with her parents [Participant].

Training Phase
Alignment
VOA Image-Caption Pairs

imSituated Image Event
destroying [Conflict.Attack]
Item [Target]: ship
Tool [Instrument]: bomb

Testing Phase
Multimedia News
For the rebels, bravado goes hand-in-hand with the desperate resistance the insurgents have mounted....

Cross-media Structured Common Representation Encoder

Cross-media Shared Event Classifier
Contact.Meet Conflict.Attack Conflict.Attack

Cross-media Shared Argument Classifier
Weakly Aligned Structured Embedding (WASE)
-- Training and Testing Phase (Cross-media shared classifiers)

Training Phase

- ACE Text Event: Liana Owen [Participant] drove from Pennsylvania to attend [Contact.Meet] the rally in Manhattan with her parents [Participant].
- Alignment: VOA Image-Caption Pairs
- imSitu Image Event: destroying [Conflict.Attack]
  - Item [Target]: ship
  - Tool [Instrument]: bomb

Testing Phase

- Multimedia News: For the rebels, bravado goes hand-in-hand with the desperate resistance the insurgents have mounted.....
- Cross-media Structured Common Representation Encoder
- Cross-media Shared Event Classifier
  - Contact.Meet
  - Conflict.Attack

Cross-media Shared Argument Classifier
- Contact.Meet
- Conflict.Attack
- Conflict.Attack
- Conflict.Attack
- Conflict.Attack
- Conflict.Attack
Weakly Aligned Structured Embedding (WASE)

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

ACE Text Event
Liana Owen [Participant] drove from Pennsylvania to attend [Contact.Meet] the rally in Manhattan with her parents [Participant].

Alignment
VOA Image-Caption Pairs

imSitu Image Event
destroying [Conflict.Attack]
Item [Target]: ship
Tool [Instrument]: bomb

Cross-media Structured Common Representation Encoder

Cross-media Shared Event Classifier
Contact.Meet
Conflict.Attack

Cross-media Shared Argument Classifier
Contact.Meet Participant
Conflict.Attack Instrument

Testing Phase
Multimedia News
For the rebels, bravado goes hand-in-hand with the desperate resistance the insurgents have mounted.....
Weakly Aligned Structured Embedding (WASE)
-- Training and Testing Phase (Cross-media shared classifiers)

**Training Phase**

**Alignment**
- VOA Image-Caption Pairs

**Weak Text Event**
Liana Owen [Participant] drove from Pennsylvania to attend [Contact.Meet] the rally in Manhattan with her parents [Participant].

**imSitu Image Event**
- Image: @ship
- Tool: bomb
- Target: Conflict.Attack

**Cross-media Structured Common Representation Encoder**

**Testing Phase**

**Multimedia News**
For the rebels, bravado goes hand-in-hand with the desperate resistance the insurgents have mounted....

**Cross-media Shared Event Classifier**
- Contact.Meet
- Conflict.Attack

**Cross-media Shared Argument Classifier**
- Contact.Meet Participant
- Conflict.Attack Instrument
Weakly Aligned Structured Embedding (WASE)
-- Training and Testing Phase (Cross-media shared classifiers)

Training Phase

**ACE Text Event**
Liana Owen [Participant] drove from Pennsylvania to attend [Contact.Meet] the rally in Manhattan with her parents [Participant].

**Alignment**
VOA Image-Caption Pairs

**imSitu Image Event**

*destroying* [Conflict.Attack]
*Item [Target]: ship*
*Tool [Instrument]: bomb*

Testing Phase

**Multimedia News**
For the rebels, bravado goes hand-in-hand with the desperate resistance the insurgents have mounted....

Cross-media Structured Common Representation Encoder

Cross-media Shared Event Classifier

Cross-media Shared Argument Classifier

Contact.Meet
Participant
Conflict.Attack
Instrument
Conflict.Attack
Attacker
Instrument
Weakly Aligned Structured Embedding (WASE)

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

**Training Phase**

- **ACE Text Event**
  - Liana Owen [Participant]
  - drove from Pennsylvania to attend [Contact.Meet] the rally in Manhattan with her parents [Participant].

- **Alignment**
  - VOA Image-Caption Pairs

- **imSitu Image Event**
  - destroying [Conflict.Attack]
  - Item [Target]: ship
  - Tool [Instrument]: bomb

**Testing Phase**

- **Multimedia News**
  - For the rebels, bravado goes hand-in-hand with the desperate resistance the insurgents have mounted.....

**Cross-media Structured Common Representation Encoder**

- **entity**
  - Liana Owen

- **region**

- **trigger**
  - attend

- **image**
  - resistance

**Cross-media Shared Event Classifier**

- Contact.Meet
- Conflict.Attack

**Cross-media Shared Argument Classifier**

- Conflict.Attack
- Participant
- Instrument
Weakly Aligned Structured Embedding (WASE)
-- System Diagram
# Experiment Results

<table>
<thead>
<tr>
<th>Training</th>
<th>Model</th>
<th>Text-Only Evaluation</th>
<th>Image-Only Evaluation</th>
<th>Multimedia Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Event Mention</td>
<td>Argument Role</td>
<td>Event Mention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
</tr>
<tr>
<td></td>
<td>JMEE</td>
<td>42.5</td>
<td>58.2</td>
<td>48.7</td>
</tr>
<tr>
<td></td>
<td>GAIL</td>
<td>43.4</td>
<td>53.5</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td>WASE$^\text{II}$</td>
<td>42.3</td>
<td>58.4</td>
<td>48.2</td>
</tr>
<tr>
<td></td>
<td>WASE$^\text{II}_{\text{att}}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WASE$^\text{II}_{\text{obj}}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VSE-C</td>
<td>33.5</td>
<td>47.8</td>
<td>39.4</td>
</tr>
<tr>
<td></td>
<td>Flat$_{\text{att}}$</td>
<td>34.2</td>
<td>63.2</td>
<td>44.4</td>
</tr>
<tr>
<td></td>
<td>Flat$_{\text{obj}}$</td>
<td>38.3</td>
<td>57.9</td>
<td>46.1</td>
</tr>
<tr>
<td></td>
<td>WASE$_{\text{att}}$</td>
<td>37.6</td>
<td>66.8</td>
<td>48.1</td>
</tr>
<tr>
<td></td>
<td>WASE$_{\text{obj}}$</td>
<td>42.8</td>
<td>61.9</td>
<td>50.6</td>
</tr>
<tr>
<td></td>
<td>43.0</td>
<td>62.1</td>
<td>50.8</td>
<td>19.5</td>
</tr>
</tbody>
</table>
## Experiment Results

<table>
<thead>
<tr>
<th>Training</th>
<th>Model</th>
<th>Text-Only Evaluation</th>
<th>Image-Only Evaluation</th>
<th>Multimedia Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Event Mention</td>
<td>Argument Role</td>
<td>Event Mention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
</tr>
<tr>
<td>Text</td>
<td>JMEE</td>
<td>42.5</td>
<td>58.2</td>
<td>48.7</td>
</tr>
<tr>
<td></td>
<td>GAIL</td>
<td>43.4</td>
<td>53.5</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td>WASE$^\text{II}$</td>
<td>42.3</td>
<td>58.4</td>
<td>48.2</td>
</tr>
<tr>
<td>Image</td>
<td>WASE$^\text{II}_\text{att}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WASE$^\text{II}_\text{obj}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Multimedia</td>
<td>VSE-C</td>
<td>33.5</td>
<td>47.8</td>
<td>39.4</td>
</tr>
<tr>
<td></td>
<td>Flat$\text{att}$</td>
<td>34.2</td>
<td>63.2</td>
<td>44.4</td>
</tr>
<tr>
<td></td>
<td>Flat$\text{obj}$</td>
<td>38.3</td>
<td>57.9</td>
<td>46.1</td>
</tr>
<tr>
<td></td>
<td>WASE$\text{att}$</td>
<td>37.6</td>
<td>66.8</td>
<td>48.1</td>
</tr>
<tr>
<td></td>
<td>WASE$\text{obj}$</td>
<td>42.8</td>
<td>61.9</td>
<td>50.6</td>
</tr>
</tbody>
</table>

The red box highlights the Multimedia Evaluation for the VSE-C model with the highest $F_1$ score of 50.6.
## Experiment Results

<table>
<thead>
<tr>
<th>Training</th>
<th>Model</th>
<th>Text-Only Evaluation</th>
<th>Image-Only Evaluation</th>
<th>Multimedia Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Event Mention</td>
<td>Argument Role</td>
<td>Event Mention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$P$</td>
<td>$R$</td>
<td>$F_1$</td>
</tr>
<tr>
<td>Text</td>
<td>JMCEE</td>
<td>42.5</td>
<td>58.2</td>
<td>48.7</td>
</tr>
<tr>
<td></td>
<td>GAIL</td>
<td>43.4</td>
<td>53.5</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td>WASE$^H$</td>
<td>42.3</td>
<td>58.4</td>
<td>48.2</td>
</tr>
<tr>
<td>Multimedia</td>
<td>WASE$^H_{att}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WASE$^H_{obj}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Image</td>
<td>VSE-C</td>
<td>33.5</td>
<td>47.8</td>
<td>39.4</td>
</tr>
<tr>
<td></td>
<td>Flat$^att$</td>
<td>34.2</td>
<td>63.2</td>
<td>44.4</td>
</tr>
<tr>
<td></td>
<td>Flat$^obj$</td>
<td>38.3</td>
<td>57.9</td>
<td>46.1</td>
</tr>
<tr>
<td>Multimedia</td>
<td>WASE$^{att}$</td>
<td>37.6</td>
<td>66.8</td>
<td>48.1</td>
</tr>
<tr>
<td></td>
<td>WASE$^{obj}$</td>
<td>42.8</td>
<td>61.9</td>
<td>50.6</td>
</tr>
</tbody>
</table>

The highest scores are highlighted in red.
# Cross-Media Coreference Accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>$P$ (%)</th>
<th>$R$ (%)</th>
<th>$F_1$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rule_based</td>
<td>10.1</td>
<td>100</td>
<td>18.2</td>
</tr>
<tr>
<td>VSE</td>
<td>31.2</td>
<td>74.5</td>
<td>44.0</td>
</tr>
<tr>
<td>Flat_att</td>
<td>33.1</td>
<td>73.5</td>
<td>45.6</td>
</tr>
<tr>
<td>Flat_obj</td>
<td>34.3</td>
<td>76.4</td>
<td>47.3</td>
</tr>
<tr>
<td>WASE_att</td>
<td>39.5</td>
<td>73.5</td>
<td>51.5</td>
</tr>
<tr>
<td>WASE_obj</td>
<td>40.1</td>
<td>75.4</td>
<td>52.4</td>
</tr>
</tbody>
</table>
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.

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

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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Event Type</th>
<th>Argument Role</th>
<th>Agent/Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>Justice.ArrestJail</td>
<td>Agent = man</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>Justice.ArrestJail</td>
<td>Entity = man</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Event Type</th>
<th>Argument Role</th>
<th>Artifact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>Movement.Transport</td>
<td>Artifact = none</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>Movement.Transport</td>
<td>Artifact = man</td>
<td></td>
</tr>
</tbody>
</table>
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.
• 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.
THANK YOU