Multimodal Knowledge Graphs
Generation Methods, Applications, and Challenges

Shih-Fu Chang

Alireza Zareian, Hassan Akbari, Brian Chen, Svebor Karaman,
Zhecan James Wang, and Haoxuan You
Columbia University

Prof. Heng Ji,
Manling Li, Di Lu, and Spencer Whitehead
University of Illinois, Urbana-Champaign
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

- Entities, events, relations, etc.
- Events describe what happens
  - Entities are characterized by the argument role they play in events

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.
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 Beyond Text

• We communicate through **multimedia**

• Our experiment shows 34% of news images contain event arguments that are not mentioned in text

*TransportPerson_Instrument* = stretcher
Why Multimodal?

- Visual data contains complementary data 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 ...
Challenges & Applications

- Challenges:
  - Parsing images/videos to structures
  - Grounding event/entities across modalities
  - Extracting complementary multimodal arguments

Diagram:
- Text IE:
  - Text graph
- Visual IE:
  - Scene graph
- Multi-Modal Knowledge Graph
- Application
Challenge 1: Parsing Images to Scene Graphs

- Extract structured representation of a scene
  - Entities and their semantic relationships
Parsing Images to Scene Graphs

- **Existing method**
  - Extract object proposals
  - Contextualize features by RNN (or message passing)
  - Classify all nodes and pairs of nodes

- **Limitations**
  - Computationally exhaustive
    - $O(n^2)$ for $n \approx 100$ 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. CVPR19)
  - Generalized formulation of scene graph generation
    - Entity-centric → bipartite representation of predicates & entities
    - Reduce computational complexity from $O(n^2)$ to sub-quadratic
    - Model argument role relations beyond (subject, object), (agent, patient) relations
Reformulate as an Event-Centric Problem

- Our work: **Visual Semantic Parsing Network (Zareian et al. CVPR20)**
  - Generalized formulation of scene graph generation
    - Entity-centric → bipartite representation of predicates & entities
    - Reduce computational complexity from $O(n^2)$ to sub-quadratic
    - Model argument role relations beyond (subject, object), (agent, patient) relations
Bipartite Embeddings for Entity & Predicate

\[ H_e^{(0)}[1] \]
\[ H_e^{(0)}[2] \]
\[ H_e^{(0)}[3] \]
\[ \cdots \]
\[ H_e^{(0)}[n_e] \]

\[ H_p^{(0)}[1] \]
\[ H_p^{(0)}[2] \]
\[ \cdots \]
\[ H_p^{(0)}[n_p] \]

Trainable Predicate Embedding Bank
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 $\rightarrow$ Roles $\rightarrow$ Predicates

Message Passing

$$H_e^{(0)}[1] \rightarrow \text{FC}_{MP\_send}^{e\rightarrow p}$$

$$H_e^{(0)}[2] \rightarrow \text{FC}_{MP\_send}^{e\rightarrow p}$$

$$H_e^{(0)}[3] \rightarrow \text{FC}_{MP\_send}^{e\rightarrow p}$$

$$H_e^{(0)}[n_e] \rightarrow \text{FC}_{MP\_send}^{e\rightarrow p}$$

$$\text{FC}_{MP\_pool}^{e\rightarrow p} \rightarrow H_p^{(1)}[1]$$

$$\text{FC}_{MP\_pool}^{e\rightarrow p} \rightarrow H_p^{(0)}[2]$$

$$\text{FC}_{MP\_pool}^{e\rightarrow p} \rightarrow H_p^{(0)}[n_p]$$
Role-Dependent Message Passing

- Bi-directional Message passing
- Entities $\leftarrow$ Roles $\leftarrow$ Predicates

$$H_e^{(1)}[1] \quad FC_{MP\_receive} \quad + \quad FC_{MP\_pool} \quad FC_{MP\_send} \quad H_p^{(1)}[1]$$

$$H_e^{(0)}[2] \quad FC_{MP\_receive} \quad + \quad FC_{MP\_pool} \quad FC_{MP\_send} \quad H_p^{(1)}[2]$$

$$H_e^{(0)}[3] \quad FC_{MP\_receive} \quad + \quad FC_{MP\_pool} \quad FC_{MP\_send} \quad H_p^{(1)}[3]$$

$$\ldots$$

$$H_e^{(0)}[n_e] \quad FC_{MP\_pool} \quad FC_{MP\_pool} \quad FC_{MP\_pool} \quad H_p^{(1)}[n_p]$$
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
Incorporate External KB (Zareian, et al, ECCV20)

- Link concepts in scene graphs to external knowledge bases such as ConceptNet
- Pass messages over bridges between scene graphs and external graphs
- Refine bridges between graphs

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>GC</th>
<th>IMP+</th>
<th>FREQ</th>
<th>SMN</th>
<th>KERN</th>
<th>GB-NET</th>
<th>GB-NET-β</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGGEN</td>
<td>mR@50</td>
<td>Y</td>
<td>3.8</td>
<td>4.3</td>
<td>5.3</td>
<td>6.4</td>
<td>6.1</td>
<td>7.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>5.4</td>
<td>5.9</td>
<td>9.3</td>
<td>11.7</td>
<td>9.8</td>
<td>11.7</td>
</tr>
<tr>
<td></td>
<td>mR@100</td>
<td>Y</td>
<td>4.8</td>
<td>5.6</td>
<td>6.1</td>
<td>7.3</td>
<td>7.3</td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>8.0</td>
<td>8.9</td>
<td>12.9</td>
<td>16.0</td>
<td>14.0</td>
<td>16.6</td>
</tr>
<tr>
<td></td>
<td>R@50</td>
<td>Y</td>
<td>20.7</td>
<td>23.5</td>
<td>27.2</td>
<td>27.1</td>
<td>26.4</td>
<td>26.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>22.0</td>
<td>25.3</td>
<td>30.5</td>
<td>30.9</td>
<td>29.4</td>
<td>29.3</td>
</tr>
<tr>
<td></td>
<td>R@100</td>
<td>Y</td>
<td>24.5</td>
<td>27.6</td>
<td>30.3</td>
<td>29.8</td>
<td>30.0</td>
<td>29.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>27.4</td>
<td>30.9</td>
<td>35.8</td>
<td>35.8</td>
<td>35.1</td>
<td>35.0</td>
</tr>
</tbody>
</table>
Scene Graph Examples of GB-NET

Ours (GB-Net)  Baseline (KERN)

Ours (GB-Net)  Baseline (KERN)
Challenge 2: Text-Visual Grounding (Akbari et al CVPR19)

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

- Challenges
  - Sensitive to domain variations
  - Abstract concept not groundable
Challenge 3: Multimodal Event & Argument Extraction

- Challenges:
  - Parsing images/videos to structures
  - Grounding entities across modalities
  - Joint extraction of multimodal argument

Diagram:
- Text IE
- Visual IE
- Text graph
- Scene graph
- Multi-Modal Knowledge Graph
- Application
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.
A New Task: Multimedia Event Extraction ($M^2E^2$)

**Input: News article text and image**

In March, Turkish forces escalated attacks on the YPG in northern Syria, forcing U.S. to deploy 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** mounted by **U.S.-led coalition forces** and Syrian warplanes. Local anti-IS activists say the air raids fail to distinguish between military and non-military targets …

**Output: Image-related Events & Visual Argument Roles**

<table>
<thead>
<tr>
<th>Event</th>
<th>Conflict.Attack</th>
<th>airstrikes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arguments</td>
<td>Attacker</td>
<td>U.S.-led coalition forces</td>
</tr>
<tr>
<td>Target</td>
<td>Target</td>
<td>airplane</td>
</tr>
<tr>
<td>Target</td>
<td>Target</td>
<td>vehicle</td>
</tr>
</tbody>
</table>
Cross-media Structured Common Space

- Treat image as another language
- Represent it with a structure that is similar to AMR in text
- Can we find a common representation?

<table>
<thead>
<tr>
<th>Linguistic Structure (Abstract Meaning Representation (AMR) / Dependency Tree)</th>
<th>Visual Semantic Graph [Zareian et al. CVPR20]</th>
</tr>
</thead>
</table>

Thai opposition protesters *attack* a bus carrying pro-government Red Shirt supporters on their way to a rally at a stadium in *Bangkok*.
Image to Event Graph

- ImSitu dataset: situation recognition (Yatskar et al., 2016)
  - Classify an image as one of 500+ FrameNet verbs (sharing part of ACE)
  - Identify 192 generic semantic roles
Weakly Aligned Structured Embedding (WASE) -- Cross-media shared representation and classifiers (Li, Zareian, et al, ACL20)
Prior work aligns image-caption vectors by triplet loss.

- We want to align two graphs, not just single vectors.

Use image-caption data for graph alignment
Use image-caption data for graph alignment

- Prior work aligns image-caption vectors by triplet loss.
- We want to align two graphs, not just single vectors.
A New Multimodal Dataset for M2E2 Evaluation
(Li, Zareian, et al, ACL20)

- Ontology: shared between ACE and imSitu
  - **Event Types**: cover 52% of ACE event types
  - **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, Time</td>
</tr>
<tr>
<td>Transaction.TransferMoney</td>
<td>Giver, Recipient, Beneficiary, Money, <strong>Instrument</strong>, Place, Time</td>
</tr>
<tr>
<td>Conflict.Attack</td>
<td>Attacker, Instrument, Place, Target, Time</td>
</tr>
<tr>
<td>Conflict.Demonstrate</td>
<td>Demonstrator, <strong>Instrument</strong>, <strong>Police</strong>, Place, Time</td>
</tr>
<tr>
<td>Contact.Phone-Write</td>
<td>Participant, <strong>Instrument</strong>, Place, Time</td>
</tr>
<tr>
<td>Contact.Meet</td>
<td>Participant, Place, Time</td>
</tr>
<tr>
<td>Justice.Arrest.Jail</td>
<td>Agent, Person, <strong>Instrument</strong>, Place, Time</td>
</tr>
<tr>
<td>Movement.Transport</td>
<td>Agent, Artifact/Person, Instrument, Destination, Origin, Time</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$_{ver}^T$</td>
<td>42.3</td>
<td>58.4</td>
<td>48.2</td>
</tr>
<tr>
<td>Image</td>
<td>WASE$_{att}^T$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WASE$_{obj}^T$</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Multimedia</td>
<td>VSE</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></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><strong>50.6</strong></td>
</tr>
</tbody>
</table>

**Multimodal Task**
Compare to Single Modality Extraction

- Image helps textual event extraction, and surrounding sentence helps visual event extraction

Iraqi security forces search [Justice.Arrest] a civilian in the city of Mosul. People celebrate Supreme Court ruling on Same Sex Marriage in front of the Supreme Court in Washington.
Application 1: Visual Commonsense Reasoning (VCR)

- Understand semantics in images and language, explore commonsense
- Provide to-the-point answer

2. What is [person1] going to do next?

- a) [person1] is going to put his hand in his pocket. 23.9%
- b) He is going to throw the paper on the ground and rant and rave at [person3] and [person2]. 11.2%
- c) [person2] is going to decide to start following a person who is out of camera’s range but in his view. 0.0%
- d) He’s going to put the fishbowl in the helicopter. 64.8%
Combine Visual Scene Graphs with VCR

- Expand input to include objects and predicate relations in graph.
- Attention transformers limited to sparse connections in scene graphs.

Graph-based Global-Local Attention Transformers (GLAT, ECCV’20)
Graph-based Global-Local Attention Transformers (Zareian, et al ECCV20)
Scene Graph + Query-Adaptive Concept Selection
- For each question, select most relevant nodes on the scene graph

<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
<th>(Entity #, Predicate #)</th>
<th>Q -&gt; A</th>
</tr>
</thead>
<tbody>
<tr>
<td>LXMERT</td>
<td>Initial Graph</td>
<td>(36,18)</td>
<td>65.09 (baseline)</td>
</tr>
<tr>
<td></td>
<td>Relevance Sel.</td>
<td>(8, x)</td>
<td>74.04 (+8.95)</td>
</tr>
<tr>
<td>GLAT (LXMERT)</td>
<td>Initial Graph</td>
<td>(36, 18)</td>
<td>65.24 (baseline)</td>
</tr>
<tr>
<td></td>
<td>Relevance Sel.</td>
<td>(26, x)</td>
<td>69.57 (+4.33)</td>
</tr>
<tr>
<td></td>
<td>Relevance Sel.</td>
<td>(18, x)</td>
<td>72.33 (+7.09)</td>
</tr>
<tr>
<td></td>
<td>Relevance Sel.</td>
<td>(8, x)</td>
<td>74.45 (+9.21)</td>
</tr>
</tbody>
</table>
Q: Why is sheep near the construction?
A: Sheep is near its natural habitat as well.
Application 2: Multimodal KG Extraction from COVID-19 Medical Papers

Figure 1.
FDA approved drugs of most interest for repurposing as potential Ebola virus treatments.

PDF images extraction, segmentation, and recognition

KG from caption text

FDA approve repaint Ebola

Multimedia Knowledge Graph Construction
Conclusions

- Multimodal Knowledge Graphs
  - Understanding semantic structures in both language and vision
  - Joint representation and models

- Applications
  - Reasoning (VCR)
  - Discovery (COVID-19)

- Challenges
  - Open-vocabulary and Self-Supervised models
  - Knowledge graphs for video
  - Commonsense Extraction from MM KG physics, behavior, causal/temporal
References


