Aligning Images and Text with Semantic Role Labels for Fine-Grained Cross-Modal Understanding

Abhidip Bhattacharyya, Cecilia Mauceri, Martha Palmer, Christoffer Heckman

firstname.lastname@colorado.edu

LREC, June 20, 2022
Introduction

1. Cross-modal retrieval
2. Lack of semantics in Vision
3. Semantic Role Labeling As Cue

Approach

1. Semantic Role aware Cross-modal Retrieval
2. Architecture

Experiments

1. Data Preparation
2. Results
3. Ablation Study
4. Fine grained Retrieval
5. Reasonable Mismatch
6. Transformers

Conclusion
Cross-modal retrieval

Lack of semantics in Vision

Semantic Role Labeling As Cue

Two men wearing khaki pants are looking at a tree that has just been felled by a saw.
Cross-modal Retrieval

Image-text database

Two men wearing khaki pants are looking at a tree that has just been felled by a saw.
How - Cross Modal Retrieval

(Lee et al., 2018; Liu et al., 2019; Li et al., 2019; Wang et al., 2020b)

1. Two branched
   - Each branch dedicated to learning representation for one modality
2. Attention: To learn correspondence
3. A loss function to embed related pair nearby
   - Sum of Negatives
   - Hard negative

Two men wearing khaki pants are looking at a tree that has just been felled by a saw.

Image encoder

Text Encoder

Ranking Loss

Image to text

Text to image

Attention module
1 Elderly man with cane bends down to look at some plants and is steadied from behind.
1. Elderly man with cane bends down to look at some plants and is steadied from behind.

2. A man and a woman are standing behind an elderly man who is looking at a bush.
Elderly man with cane bends down to look at some plants and is steadied from behind.

A man and a woman are standing behind an elderly man who is looking at a bush.

A man holds up an older man as the older man bends down to check out plants.
Semantics?

1. Elderly man with cane bends down to look at some plants and is steadied from behind.
2. A man and a woman are standing behind an elderly man who is looking at a bush.
3. A man holds up an older man as the older man bends down to check out plants.
4. An older man in a white short-sleeve shirt admiring a bush.
5. Elderly man with a cane bends over near a man and woman.
Role Aware REtrieval system

- Semantic Role as cue for retrieval
Role Aware REtrieval system

The man is aiming to shoot something while his dog watches.

A man aiming a rifle with a dog standing beside him.
Semantic Role Labeling

Semantic Role

- Captures – ‘who’ is doing ‘what’ to ‘whom’ ‘where’, ‘when’ and ‘how’?
Semantic Role

Captures – ‘who’ is doing ‘what’ to ‘whom’ ‘where’, ‘when’ and ‘how’?

Carl gave food to his pet.

Carl gave his pet food.

The food was given to his pet by Carl.
Semantic Role Labeling

Semantic Role

Captures – ‘who’ is doing ‘what’ to ‘whom’ ‘where’, ‘when’ and ‘how’?

- *Carl* gave *food* to *his pet*.

- *Food* was given to *his pet* by *Carl*.
  - what: *Food*, did: *was given*, whom: *his pet*, who: *Carl*.
Semantic Role Labeling

Semantic Role

- Captures – ‘who’ is doing ‘what’ to ‘whom’ ‘where’, ‘when’ and ‘how’?

<table>
<thead>
<tr>
<th>Arg0</th>
<th>prototypical agent</th>
<th>Arg3</th>
<th>starting point, benefactive, attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg1</td>
<td>prototypical patient</td>
<td>Arg4</td>
<td>ending point</td>
</tr>
<tr>
<td>Arg2</td>
<td>instrument, benefactive, attribute</td>
<td>ArgM</td>
<td>modifier</td>
</tr>
</tbody>
</table>

**Figure**: Semantic Roles Presented in PropBank (Palmer et al., 2005)
Semantic Role Labeling

Semantic Role
- Captures – ‘who’ is doing ‘what’ to ‘whom’ ‘where’, ‘when’ and ‘how’?

Proposition
- The baby is playing on the porch while parents are watching.
- The baby is playing on the porch while parents are watching.
1. Introduction
   - Cross-modal retrieval
   - Lack of semantics in Vision
   - Semantic Role Labeling As Cue

2. Approach
   - Semantic Role aware Cross-modal Retrieval
   - Architecture

3. Experiments
   - Data Preparation
   - Results
   - Ablation Study
   - Fine grained Retrieval
   - Reasonable Mismatch
   - Transformers

4. Conclusion
Role Aware REtrieval system

RARE

- Semantic Role as cue for retrieval
Role Aware REtrieval system

The man is aiming to shoot something while his dog watches

A man aiming a rifle with a dog standing beside him
Role Aware REtrieval system

Semantic Role aware Cross-modal Retrieval

Architecture

[The man]_Arg0 is [aiming]_predicate [to shoot something]_Arg1 [while his dog watches]_ArgM-TMP

[A man]_Arg0 [aiming]_predicate [a rifle]_Arg1 [with a dog standing beside him]_ArgM-MNR
Role Aware REtrieval system

RARE

- Semantic Role as cue for retrieval
- Two branch approach
  - each branch will have corresponding semantic role annotation
Semantic Role aware Cross-modal Retrieval

Architecture
Architecture

[A man]_A0 aiming [a rifle]_A1 [with a dog standing beside him]_AM-MNR.
A man aiming a rifle [with a dog standing beside him].
Semantic Role aware Cross-modal Retrieval

Architecture

[A man]_**A0** aiming [a rifle]_**A1** [with a dog standing beside him]_**AM-MNR**.
Architecture (bi-directional focal attention) (Liu et al., 2019)

Pre-assign attention

\[ w_{i,j} = \sigma(\alpha \frac{u_i^T v_j}{||u_i|| ||v_j||}) \]
Architecture (bi-directional focal attention) (Liu et al., 2019)

Identify Relevant fragments

\[ F(w_{i,j}) = \sum_{t=1}^{n} |w_{i,j} - w_{i,t}| \times g(w_{i,j}) \]

\[ H(w_{i,j}) = \mathbb{1}(F(w_{i,j}) > 0) \]
Architecture (bi-directional focal attention) (Liu et al., 2019)

Resign Attention

\[ w_{i,j}' = \frac{w_{i,j}H(w_{i,j})}{\sum_{t=1}^{n} w_{i,t}H(w_{i,t})} \]
1. **Introduction**
   - Cross-modal retrieval
   - Lack of semantics in Vision
   - Semantic Role Labeling As Cue

2. **Approach**
   - Semantic Role aware Cross-modal Retrieval
   - Architecture

3. **Experiments**
   - Data Preparation
   - Results
   - Ablation Study
   - Fine grained Retrieval
   - Reasonable Mismatch
   - Transformers

4. **Conclusion**
We used Flickr 30K entity datasets (Plummer et al., 2017)
- Mapping between text entity mentions and the image bounding boxes
- Semantic role annotations
  - text descriptions are annotated with SRL system (Gung and Palmer, 2021)
  - semantic roles from text descriptions are transferred to images by entity mapping (Plummer et al., 2017)
A man shoots a rifle while a dog looks on.
RARE- Data Preparation

A man shoots a rifle while a dog looks on.

[A man]_Arg0 shoots [a rifle]_Arg1 [while a dog looks on ]_ArgM-TMP

SRL (Gung & Palmer, 2021)

A man shoots a rifle while a [dog]_Arg0 looks on.
RARE - Data Preparation

Data Preparation

[A man]_Arg0 shoots [a rifle]_Arg1 [while a dog looks on.]_ArgM-TMP

A man shoots a rifle while a dog looks on.

flickr30k entity mapping

SRL (Gung & Palmer, 2021)

A man shoots a rifle while a [dog]_Arg0 looks on.

flickr30k entity mapping
[A young lady wearing blue and black]_Arg0 is running [past an orange cone]_ArgM-DIR.

[The child in the green one piece suit]_Arg0 is walking [past a store window]_ArgM-DIR.

[A man]_Arg0 skis past another man displaying [paintings]_Arg1 [in the snow]_ArgM-LOC.
<table>
<thead>
<tr>
<th>Model</th>
<th>Text to Image</th>
<th>Image To Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R5</td>
</tr>
<tr>
<td>Wang et al. (2019)</td>
<td>50.4</td>
<td>78.7</td>
</tr>
<tr>
<td>Ren et al. (2016)</td>
<td>50.6</td>
<td>79.8</td>
</tr>
<tr>
<td>Liu et al. (2019)</td>
<td>50.8</td>
<td>78.4</td>
</tr>
<tr>
<td>Wang et al. (2020b)</td>
<td>53.5</td>
<td>79.6</td>
</tr>
<tr>
<td>Huang and Wang (2019)</td>
<td>53.8</td>
<td>79.8</td>
</tr>
<tr>
<td>Li et al. (2019)</td>
<td>54.7</td>
<td>81.8</td>
</tr>
<tr>
<td>Wang et al. (2020a)</td>
<td>52.9</td>
<td>80.4</td>
</tr>
<tr>
<td>Liu et al. (2020)</td>
<td>57.4</td>
<td>82.3</td>
</tr>
<tr>
<td>RARE (ours)</td>
<td><strong>67.8</strong></td>
<td><strong>83</strong></td>
</tr>
</tbody>
</table>
## Ablation Study

<table>
<thead>
<tr>
<th>Model Ablation</th>
<th>Text to Image</th>
<th></th>
<th>Image To Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R5</td>
<td>R10</td>
</tr>
<tr>
<td>BFAN base model</td>
<td>53.5</td>
<td>79.6</td>
<td>73.4</td>
</tr>
<tr>
<td>+ SRL encodings</td>
<td>65.1</td>
<td>79.8</td>
<td>86.9</td>
</tr>
<tr>
<td>+ GCN</td>
<td><strong>67.8</strong></td>
<td><strong>83</strong></td>
<td><strong>88.4</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input Ablation</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Image SRL only</td>
<td>40.9</td>
<td>45</td>
<td>58</td>
</tr>
<tr>
<td>Text SRL only</td>
<td>36.9</td>
<td>36.9</td>
<td>49</td>
</tr>
<tr>
<td>Both</td>
<td><strong>67.8</strong></td>
<td><strong>83</strong></td>
<td><strong>88.4</strong></td>
</tr>
</tbody>
</table>
Fine grained retrieval - Image to text

Query 2

Retrieved caption:
A man playing a musical instrument

Parsed SRLs for retrieved caption:
[A man]_Arg0 [playing]_V [a musical instrument]_Arg2
Fine grained retrieval - Image to text

Query 1

Retrieved caption:
A man with glasses is sitting in a chair playing the oboe while a man in a purple shirt plays percussion and spectators look on.

Parsed SRLs for retrieved caption:
1. [A man with glasses]_Arg0 [playing]_V [the oboe]_Arg2 [while a man in a purple shirt plays percussion and spectators look on]_ArgM-TMP
2. [A man with glasses]_Arg0 [playing]_V [the oboe]_Arg2 [while a man in a purple shirt plays percussion and spectators look on]_ArgM-TMP
3. ... [a man in a purple shirt]_Arg0 [plays]_V [percussion]_Arg1 ...

Query 2

Retrieved caption:
A man playing a musical instrument

Parsed SRLs for retrieved caption:
[A man]_Arg0 [playing]_V [a musical instrument]_Arg2
People standing on rocks by a river.
- [People].Arg0standing [on rocks by a river].ArgM-LOC.
A woman and her son sitting on top of a big rock looking tired.
- [A woman and her son]$_{Arg0}$ sitting [on top of a big rock]$_{ArgM-LOC}$ looking tired.
- [A woman and her son]$_{Arg0}$ sitting on top of a big rock looking [tired]$_{ArgM-MNR}$.
A boy ties his shoe while a woman carrying straw hats looks on atop a rock in front of a body of water.
- [A boy]_Arg0 is ties [his shoe]_Arg1[while...]_ArgM-TMP
- ... while [a woman]_Arg0 carrying [straw hats]_Arg1
- ... [a woman]_Arg0 carrying straw hats looks [on atop a rock in front of a body of water]_ArgM-LOC
**Experiments**

---

**Fine grained retrieval - Text to image**

3. People standing on rocks by a river.
   - [People] \_Arg0 \textbf{standing} [on rocks by a river] \_ArgM-LOC.

3. A woman and her son sitting on top of a big rock looking tired.
   - [A woman and her son] \_Arg0 \textbf{sitting} [on top of a big rock] \_ArgM-LOC looking tired.
   - [A woman and her son] \_Arg0 sitting on top of a big rock \textbf{looking} [tired] \_ArgM-MNR.

3. A boy ties his shoe while a woman carrying straw hats looks on atop a rock in front of a body of water.
   - [A boy] \_Arg0 \textbf{is ties} [his shoe] \_Arg1[while...] \_ArgM-TMP
   - ... while [a woman] \_Arg0 \textbf{carrying} [straw hats] \_Arg1
   - ... [a woman] \_Arg0 carrying straw hats \textbf{looks} [on atop a rock in front of a body of water] \_ArgM-LOC.
Reasonable Mismatching

Ground Truth: [A fashionable young woman seated on a bench]$_{Arg0}$ gazes [into a makeup mirror]$_{ArgM-DIR}$.
Retrieved: [An elderly man]$_{Arg0}$ sitting on [a bench]$_{Arg2}$ [while reading a book]$_{ArgM-TMP}$.

Ground Truth: [A red car]$_{Arg0}$ driving [over a bridge]$_{ArgM-LOC}$.
Retrieved: [A red car]$_{Arg0}$ travels down [the street]$_{ArgM-DIR}$.
Reasonable Mismatching

**Ground Truth:** [A little boy]_{Arg0} playing [Game-Cube]_{Arg1} [at a McDonald’s]_{ArgM-LOC}.

**Retrieved:** [The child]_{Arg0} is playing [croquette]_{Arg1} [by the truck]_{ArgM-LOC}. 

![Image of a child playing with a game cube at McDonald's](image_url)
### SRL Study

<table>
<thead>
<tr>
<th>Role</th>
<th>Description of Role</th>
<th>Dataset</th>
<th>Image to Text</th>
<th>Text To Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg0</td>
<td>object which instigates the verb</td>
<td>158969</td>
<td>4690 0.96</td>
<td>4985 0.94</td>
</tr>
<tr>
<td>Arg1</td>
<td>object which is affected by the verb</td>
<td>161841</td>
<td>4187 0.96</td>
<td>5025 0.82</td>
</tr>
<tr>
<td>Arg2</td>
<td>object which affects the verb</td>
<td>63853</td>
<td>1468 0.89</td>
<td>1967 0.72</td>
</tr>
<tr>
<td>ArgM-LOC</td>
<td>location of object or action</td>
<td>47866</td>
<td>910 0.85</td>
<td>1482 0.60</td>
</tr>
<tr>
<td>ArgM-TMP</td>
<td>describes time</td>
<td>17458</td>
<td>406 0.93</td>
<td>574 0.67</td>
</tr>
<tr>
<td>ArgM-DIR</td>
<td>direction of motion</td>
<td>18933</td>
<td>316 0.84</td>
<td>600 0.50</td>
</tr>
<tr>
<td>ArgM-MNR</td>
<td>manner of performing an action</td>
<td>15503</td>
<td>306 0.73</td>
<td>457 0.56</td>
</tr>
<tr>
<td>ArgM-PRD</td>
<td>adjunct of an action</td>
<td>3698</td>
<td>74 0.81</td>
<td>101 0.64</td>
</tr>
<tr>
<td>ArgM-PRP</td>
<td>purpose of an action</td>
<td>2999</td>
<td>58 0.85</td>
<td>108 0.48</td>
</tr>
<tr>
<td>ArgM-COM</td>
<td>who an action was done with</td>
<td>1618</td>
<td>47 0.85</td>
<td>55 0.69</td>
</tr>
<tr>
<td>Arg3</td>
<td>starting position of action</td>
<td>1705</td>
<td>32 0.81</td>
<td>47 0.53</td>
</tr>
</tbody>
</table>
### Transfomer based method

<table>
<thead>
<tr>
<th>Model</th>
<th>Text to Image</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R5</td>
<td>R10</td>
</tr>
<tr>
<td>RARE (ours)</td>
<td>67.8</td>
<td>83.0</td>
<td>88.4</td>
</tr>
<tr>
<td>Chen et al. (2020)</td>
<td>76.0</td>
<td><strong>93.4</strong></td>
<td><strong>96.7</strong></td>
</tr>
<tr>
<td>Ren et al. (2021)</td>
<td><strong>76.3</strong></td>
<td>93.3</td>
<td>95.6</td>
</tr>
</tbody>
</table>

**Table:** Comparison with transformer based approaches on flickr.
Presentation Outline

1. Introduction
   - Cross-modal retrieval
   - Lack of semantics in Vision
   - Semantic Role Labeling As Cue

2. Approach
   - Semantic Role aware Cross-modal Retrieval
   - Architecture

3. Experiments
   - Data Preparation
   - Results
   - Ablation Study
   - Fine grained Retrieval
   - Reasonable Mismatch
   - Transformers

4. Conclusion
Incorporating semantic roles in image-text retrieval improves cross-modal retrieval specifically image retrieval. It allows retrieval of varied and fine-grained results.

Limitations:
- Needs Image annotations
- ARG-M roles are hard to align
- Application of more advanced network

Future Work:
- Automatic role annotation of image bounding boxes
- Creating semantic annotation for image data


Thank you!

email to: {Abhidip.Bhattacharyya, Cecilia.Maucery, Martha.Palmer, Christoffer.Heckman}@colorado.edu