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

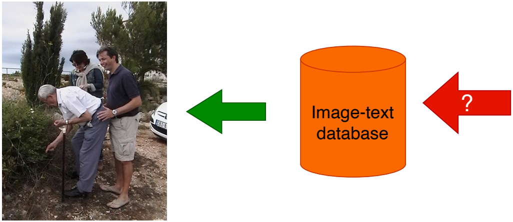
University of Colorado **Boulder**

Presentation Outline

Cross-modal Retrieval

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

Cross-modal Retrieval

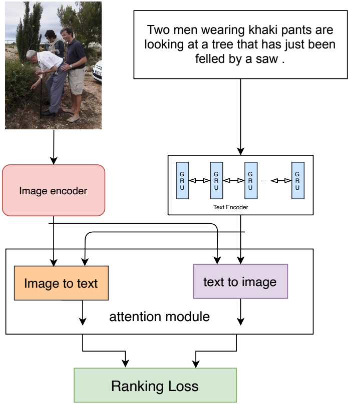


Two men wearing khaki pants are looking at a tree that has just been felled by a saw .

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



Semantics?

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

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.

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.

Semantics?

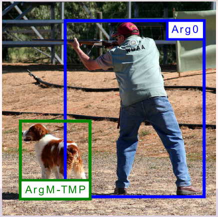
1. Elderly man with cane bends down to look at some plants and is steadied from behind.
2. A man and a woman are standing behind an elderly man who is looking at a bush.
3. A man holds up an older man as the older man bends down to check out plants.
4. An older man in a white short-sleeve shirt admiring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

Role Aware REtrieval system

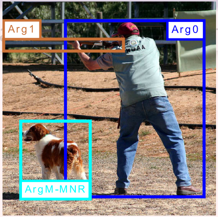
RARE

- 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

RARE

- Sem... for retrieval

Semantic Role Labeling

Semantic Role

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

Semantic Role Labeling

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'?

$\underbrace{\textit{Carl}}_{\textit{who}}\ \underbrace{\textit{gave}}_{\textit{did}}\ \underbrace{\textit{food}}_{\textit{what}}$ to $\underbrace{\textit{his pet}}_{\textit{whom}}$.

$\underbrace{\textit{Carl}}_{\textit{who}}\ \underbrace{\textit{gave}}_{\textit{did}}\ \underbrace{\textit{his pet}}_{\textit{whom}}\ \underbrace{\textit{food}}_{\textit{what}}$.

$\underbrace{\textit{Food}}_{\textit{what}}\ \underbrace{\textit{was given}}_{\textit{did}}$ to $\underbrace{\textit{his pet}}_{\textit{whom}}$ by $\underbrace{\textit{Carl}}_{\textit{who}}$.

Semantic Role Labeling

Semantic Role

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

Arg0	prototypical agent	Arg3	starting point, benefactive, attribute
Arg1	prototypical patient	Arg4	ending point
Arg2	instrument, benefactive, attribute	ArgM	modifier

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

- $\underbrace{\textit{The baby}}_{Arg0}\ \underbrace{\textit{is playing}}_{Pred}\ \underbrace{\textit{on the porch}}_{AM-LOC}\ \underbrace{\textit{while parents are watching}}_{AM-TMP}.$
- $\textit{The baby is playing on the porch while}\ \underbrace{\textit{parents}}_{Arg0}\ \underbrace{\textit{are watching}}_{Pred}.$

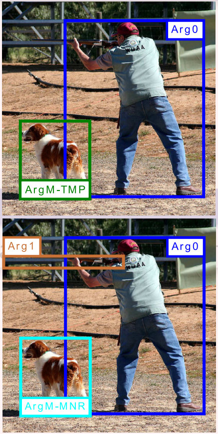
Presentation Outline

Role Aware REtrieval system

RARE

- Semantic Role as cue for retrieval

Role Aware REtrieval system



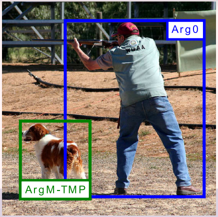
The man is aiming to shoot something while his dog watches

A man aiming a rifle with a dog standing beside him

RARE

- Sem... for retrieval

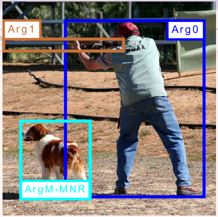
Role Aware REtrieval system



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

RARE

- Sem for retrieval



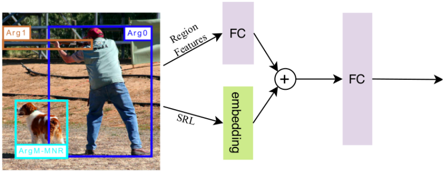
[A man]_Arg0 [aiming]_predicate [a rifle]_Arg1 [with a dog standing beside him]_ArgM-MNR

Role Aware REtrieval system

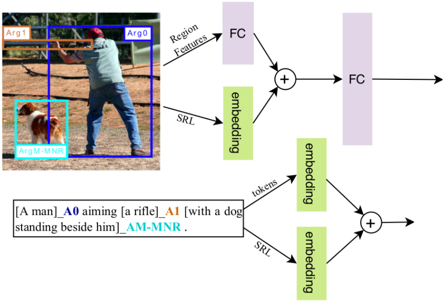
RARE

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

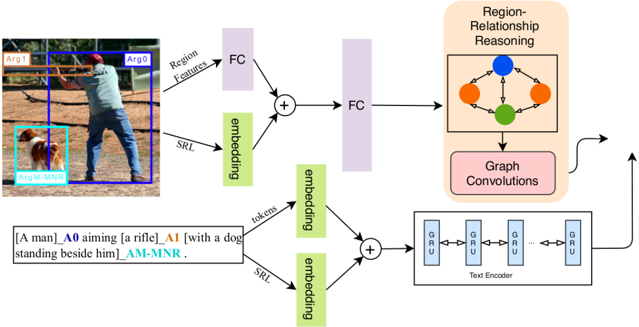
Architecture



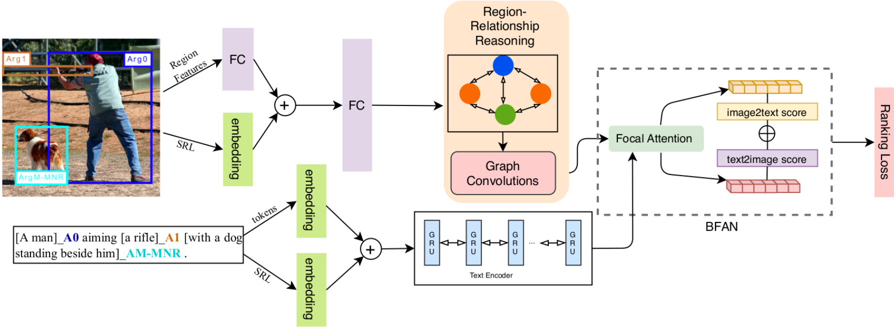
Architecture



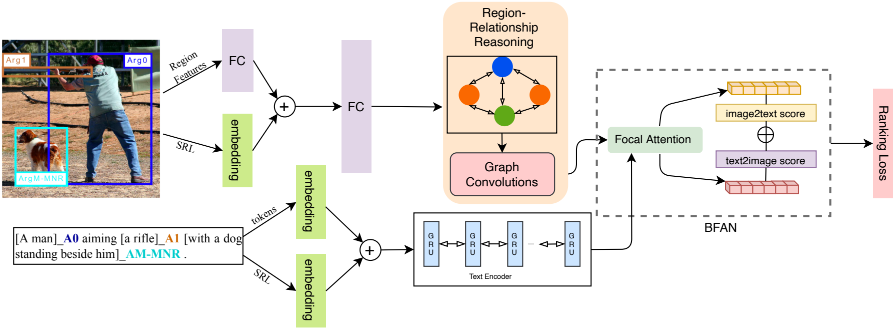
Architecture



Architecture



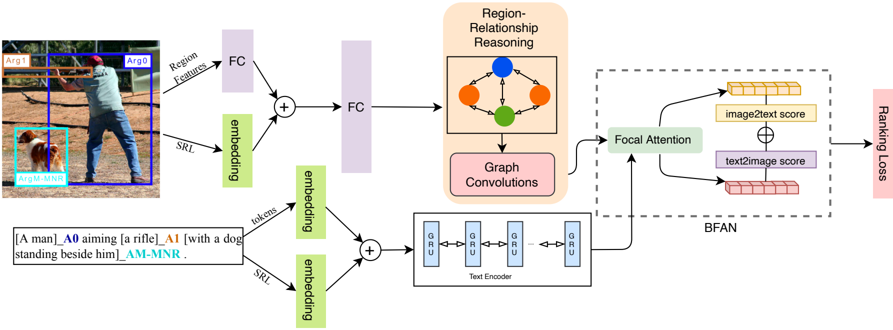
Architecture (bi-directional focal attention) (Liu et al., 2019)



Pre-assign attention

- $w_{i,j} = \sigma(\alpha \frac{u_i^T v_j}{\|u_i\|\|v_j\|})$

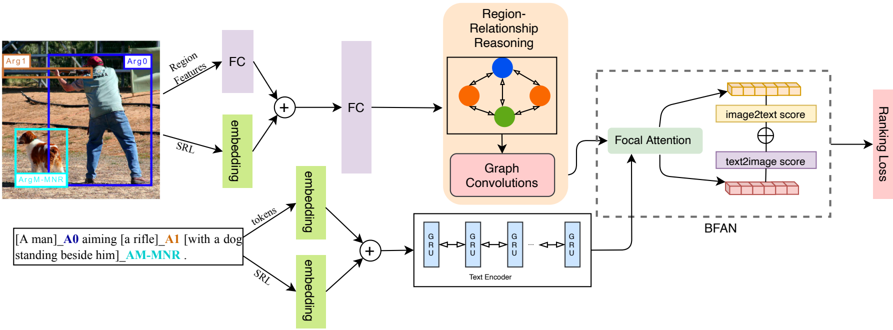
Architecture (bi-directional focal attention) (Liu et al., 2019)



Identify Relevant fragments

- $F(w_{i,j}) = \sum_{t=1}^{n} |w_{i,j} - w_{i,t}| \times g(w_{i,j})$
- $H(w_{i,j}) = \mathbb{I}(F(w_{i,j}) > 0)$

Architecture (bi-directional focal attention) (Liu et al., 2019)



Resign Attention

- $$w'_{i,j} = \frac{w_{i,j} H(w_{i,j})}{\sum_{t=1}^{n} w_{i,t} H(w_{i,t})}$$

Presentation Outline

RARE- Data Preparation

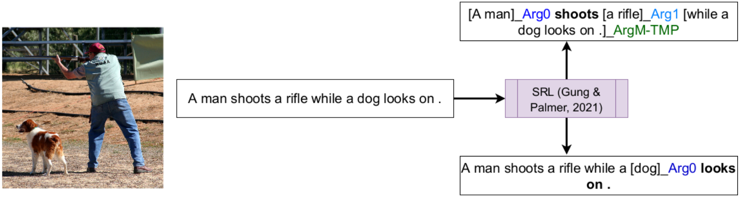
RARE

- 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)

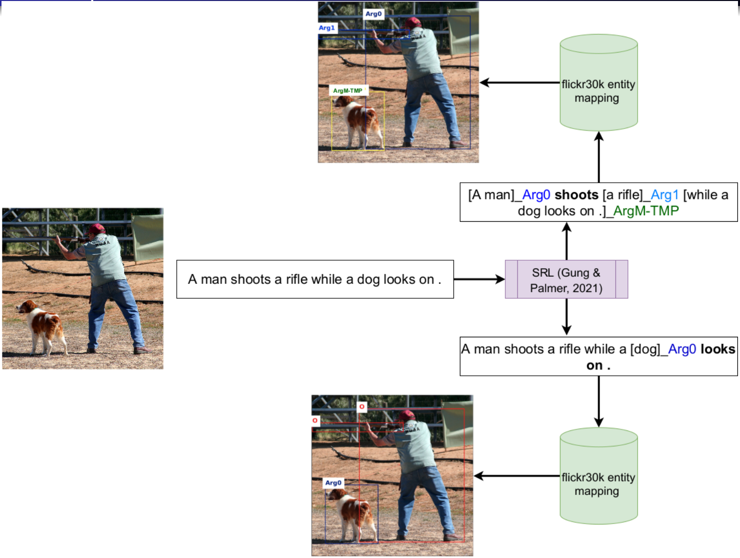
RARE- Data Preparation

A man shoots a rifle while a dog looks on .

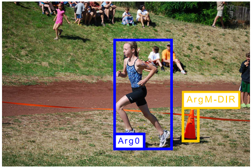
RARE- Data Preparation



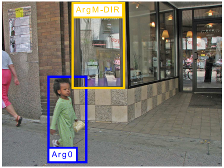
RARE- Data Preparation



Experiments



[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.

Experiments

Comparative Study

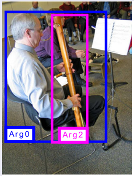
Model	Text to Image			Image To Text		
	R1	**R5**	**R10**	**R1**	**R5**	**R10**
Wang et al. (2019)	50.4	78.7	86.1	70	91.8	95.1
Ren et al. (2016)	50.6	79.8	87.6	68.5	90.9	95.5
Liu et al. (2019)	50.8	78.4	-	68.1	91.4	-
Wang et al. (2020b)	53.5	79.6	86.8	71.8	91.7	95.5
Huang and Wang (2019)	53.8	79.8	-	**85.2**	**96.7**	-
Li et al. (2019)	54.7	81.8	88.2	71.3	90.6	96
Wang et al. (2020a)	52.9	80.4	87.8	73.5	92.1	95.8
Liu et al. (2020)	57.4	82.3	**89.0**	76.4	94.3	97.3
RARE (ours)	**67.8**	**83**	88.4	76.3	93.4	96.6

Ablation Study

Model	Text to Image			Image To Text		
	R1	R5	R10	R1	R5	R10
Model Ablation						
BFAN base model	53.5	79.6	73.4	72.6	93	96
+ SRL encodings	65.1	79.8	86.9	74.2	93.1	96.5
+ GCN	**67.8**	**83**	**88.4**	**76.3**	**93.4**	**96.6**
Input Ablation						
Image SRL only	40.9	45	58	43.8	76.5	86.3
Text SRL only	36.9	36.9	49	40.8	69.6	80.8
Both	**67.8**	**83**	**88.4**	**76.3**	**93.4**	**96.6**

Fine grained retrieval- Image to text

Query 2

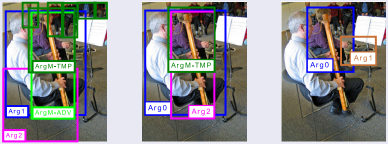


Retrieved caption:
A man playing a musical instrument

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.

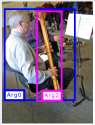
Parsed SRLs for retrieved caption:

1. [A man with glasses]_**Arg0** is [sitting]_V in [a chair]_**Arg2** [playing the oboe]_**ArgM-ADV** [while a man in a purple shirt plays percussion and spectators look on]_**ArgM-TMP**

2. [A man with glasses]_**Arg0** is sitting in a chair [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



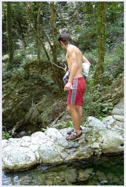
Retrieved caption:
A man playing a musical instrument

Parsed SRLs for retrieved caption:
[A man]_**Arg0** [playing]_V [a musical instrument]_**Arg2**

Experiments

Fine grained retrieval-Text to image

1. People standing on rocks by a river .
   – [People]_Arg0standing [on rocks by a river]_ArgM-LOC.



Fine grained retrieval-Text to image

2. 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-LOClooking tired .
   - [A woman and her son]_Arg0 sitting on top of a big rock **looking** [tired]_ArgM-MNR.

Experiments

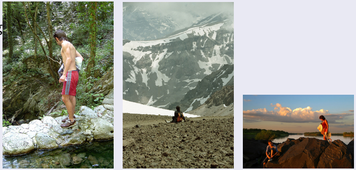
Fine grained retrieval - Text to image

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 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

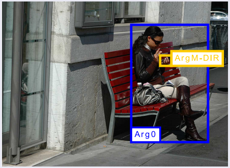
Experiments

Fine grained retrieval - Text to image

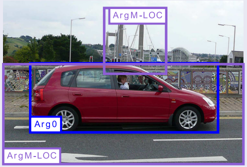
3. People standing on rocks by a river .
   - [People]_Arg0 **standing** [on rocks by a river]_ArgM-LOC.
3. A woman and her son sitting on top of a big rock looking tired .
   - [A woman and her son]_Arg0 **sitting** [on top of a big rock]_ArgM-LOClooking tired .
   - [A woman and her son]_Arg0 sitting on top of a big rock **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 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



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

Experiments

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.

SRL Study

Role	Description of Role	Dataset	Image to Text		Text To Image	
		N	N	R@1	N	R@1
Arg0	object which instigates the verb	158969	4690	0.96	4985	0.94
Arg1	object which is affected by the verb	161841	4187	0.96	5025	0.82
Arg2	object which affects the verb	63853	1468	0.89	1967	0.72
ArgM-LOC	location of object or action	47866	910	0.85	1482	0.60
ArgM-TMP	describes time	17458	406	0.93	574	0.67
ArgM-DIR	direction of motion	18933	316	0.84	600	0.50
ArgM-MNR	manner of performing an action	15503	306	0.73	457	0.56
ArgM-PRD	adjunct of an action	3698	74	0.81	101	0.64
ArgM-PRP	purpose of an action	2999	58	0.85	108	0.48
ArgM-COM	who an action was done with	1618	47	0.85	55	0.69
Arg3	starting position of action	1705	32	0.81	47	0.53

Experiments

Transfomer based method

Model	Text to Image			Image To Text		
	R1	**R5**	**R10**	**R1**	**R5**	**R10**
RARE (ours)	67.8	83.0	88.4	76.3	93.4	96.6
Chen et al. (2020)	76.0	**93.4**	**96.7**	85.8	97.8	98.8
Ren et al. (2021)	**76.3**	93.3	95.6	**88.3**	**98.6**	99.3

Table: Comparison with transformer based approaches on flickr.

Presentation Outline

Conclusions

Summary

- Incorporating semantic roles in image-text retrieval
- Improves corss-modal retrieval specifically image retrieval
- Allow retrieval of varied and fine-grained results.

Limitations

- Needs Image annotations
- ARG-M roles are hard to allign
- Application of more advanced network

Future Work

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

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020). Uniter: Universal image-text representation learning. In *ECCV*.

Gung, J. and Palmer, M. (2021). Predicate representations and polysemy in verbnet semantic parsing. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 51–62, Groningen, The Netherlands (online). Association for Computational Linguistics.

Huang, Y. and Wang, L. (2019). Acmm: Aligned cross-modal memory for few-shot image and sentence matching.

Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. (2018). Stacked cross attention for image-text matching.

Li, K., Zhang, Y., Li, K., Li, Y., and Fu, Y. (2019). Visual semantic reasoning for image-text matching.

Liu, C., Mao, Z., Liu, A.-A., Zhang, T., Wang, B., and Zhang, Y. (2019). Focus your attention: A bidirectional focal attention network for image-text matching. In *ACM International Conference on Multimedia*, page 3–11.

Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., and Zhang, Y. (2020). Graph structured network for image-text matching. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10921–10930.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2017). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vision*, 123(1):74–93.

Ren, S., Lin, J., Zhao, G., Men, R., Yang, A., Zhou, J., Sun, X., and Yang, H. (2021). Learning relation alignment for calibrated cross-modal retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 514–524, Online. Association for Computational Linguistics.

Ren, Z., Jin, H., Lin, Z., Fang, C., and Yuille, A. (2016). Joint image-text representation by gaussian visual-semantic embedding. In *ACM International Conference on Multimedia*, page 207–211.

Wang, H., Zhang, Y., Ji, Z., Pang, Y., and Ma, L. (2020a). Consensus-aware visual-semantic embedding for image-text matching. pages 18–34. Springer.

Wang, S., Wang, R., Yao, Z., Shan, S., and Chen, X. (2020b). Cross-modal scene graph matching for relationship-aware image-text retrieval.

Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., and Fan, X. (2019). Position focused attention network for image-text matching. In *International Joint Conference on Artificial Intelligence*, pages 3792–3798.

Thank you!

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