Improving the Image Discrimination Ability for CLIP-Model via Semantic Graphs through Graph Convolutional Network

Sangmyeong Lee[1], Seitaro Shinagawa[1], Koichiro Yoshino[1, 2], Satoshi Nakamura[1]

1. Nara Institute of Science and Technology 2. Guardian Robot Project, RIKEN

ABSTRACT

Contrastive Language Image Pre-training

PROBLEM : Vulnerability to Structural Ambiguity

- Possible Misinterpretation of the User's Intention

KEY IDEAS

- Fine-tuning the CLIP-Model with Semantic Graphs as Inputs
- Embedding by Graph Convolutional Networks (GCN)

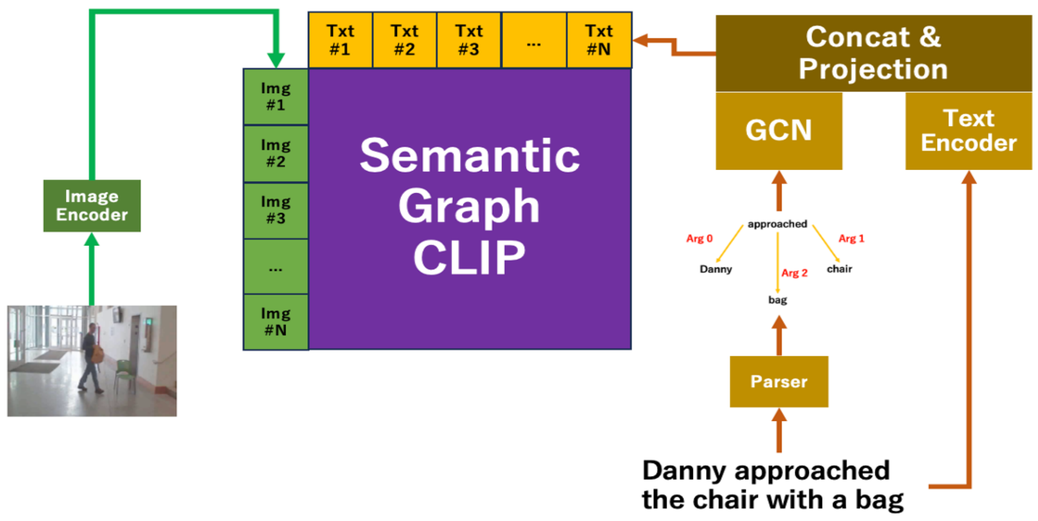
EXPERIMENTAL RESULTS

- Leveraging Semantic Graphs is Effective for Disambiguation

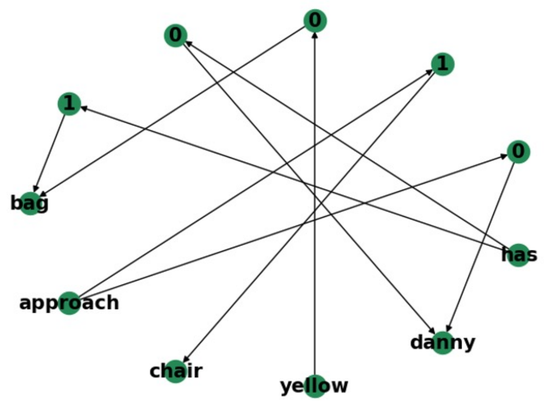
```
(S
    (NP (NNP danny))
    (VP (VBD approached)
        (NP (DT the) (JJ green) (NN chair))
        (PP (IN with) (NP (DT a) (JJ yellow) (NN bag)))
    )
)
```

```
(S
    (NP (NNP danny))
    (VP (VBD approached)
        (NP
            (NP (DT the) (JJ green) (NN chair))
            (PP (IN with) (NP (DT a) (JJ yellow) (NN bag))
            )
        )
    )
)
```

PROPOSED METHODOLOGY



- CLIP's Image Encoder Frozen
- Language Inputs : Graph Vector + Text Vector
- Graph by GCN, Text by CLIP's Text Encoder
- Combination by Concatenation and Projection



- Core Meaning without Superficial Info.
- Interconnecting Predicates with its Arguments
- Arguments' Order as Semantic Roles
- GCN to Encode Interrelationships

EXPERIMENTAL SETUPS

Does the Proposal Improve the CLIP's Performance?

DATA

- Language and Vision Ambiguity (LAVA) [1]
  2 pairs of Image-Graph for a Single Ambiguous Text
  Assess the Model's **DISAMBIGUATION ABILITY**
- Microsoft COCO
  Assess the Model's **GENERALISABILITY**

Metrics

- Discrimination Accuracy
  Accuracy of Correctly Matching Graphs with Images
- Recall@K
  K-top Retrieval Results from the whole Test Data

ANALYTICAL SETUPS

Does the Proposal Really Understand the Graph?

- Integrated Gradients [2] Method
- To Observe if the Difference in Graphs Attribute to the Model's Inference

EXPERIMENTAL RESULTS

			text-to-image			image-to-text		
Data	Model	Accuracy (%)	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
LAVA	$CLIP_{plain}$	50.0	0.264	0.708	**0.905**	0.270	0.753	0.893
	$CLIP_{tree\ as\ text}$	74.72	**0.405**	**0.753**	0.882	**0.388**	**0.803**	**0.899**
	$CLIP_{GCN}$	**77.53**	0.298	0.657	0.775	0.365	0.657	0.775
COCO	$CLIP_{plain}$	NA	**0.391**	**0.654**	**0.765**	0.408	0.680	0.787
	$CLIP_{tree\ as\ text}$	NA	0.371	0.645	0.761	**0.410**	**0.683**	**0.795**
	$CLIP_{GCN}$	NA	0.198	0.477	0.617	0.234	0.519	0.661

LAVA

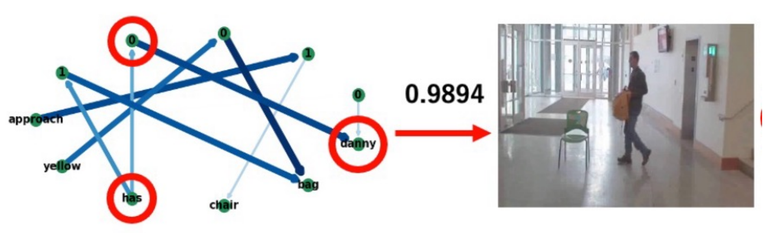
- 'Tree as Text' Showed the Best Results in R@K in an Overall Sense
- Semantic GCN Showed the Best Performance in Accuracy

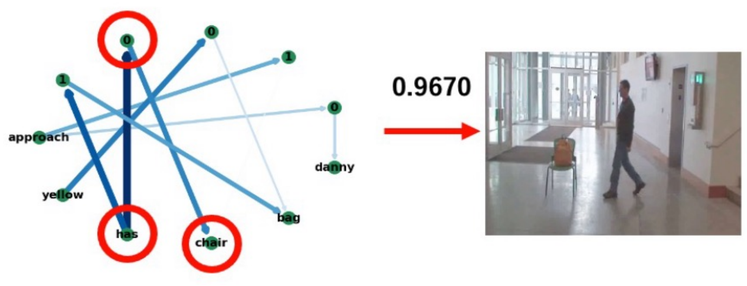
COCO

- 'Tree as Text' Compatible with the Baseline
- GCN Outnumbered
- In Generality, Strong Pre-training would be in Need

ANALYTICAL RESULTS

Danny approached the chair with a yellow bag

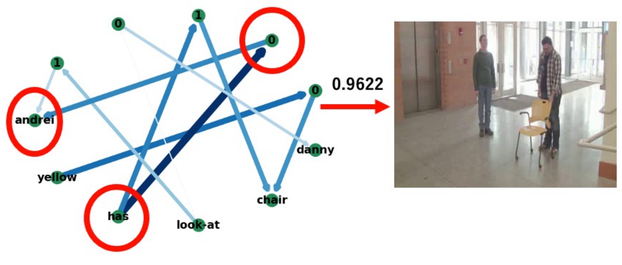




Successful Classification

- Difference in Meaning : Chair's Position
- Difference in Graphs Contributes to Inference
- With more than 2 Personnels, Inference Fails

Danny looked at Andrei having a yellow chair



Unsuccessful Classification

- With more than 2 Personnels with Pronouns, Inference Failed despite Right Attention
- The Model Became Overfitted, Adopting Strategies Irrelevant to the Actual Difference in Graphs

FUTURE CHALLENGES

- Collecting Data Free from LAVA's Weakness
- Strengthening Generalisability
- Anlysis on Multimodal

REFERENCES

[1] Y. Berzak et al., Do You See What I Mean? Visual Resolution of Linguistic Ambiguities, EMNLP, 2015.
[2] M. Sundararajan et al., Axiomatic Attribution for Deep Networks, ICMR, 2017.