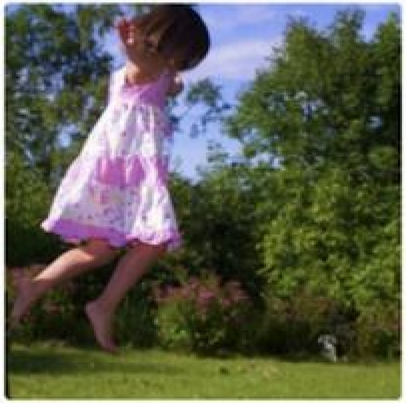
L2C: Describing Visual Differences Needs Semantic Understanding of Individuals

An Yan, Xin Eric Wang, Tsu-Jui Fu, William Yang Wang

UC San Diego, UC Santa Cruz, UC Santa Barbara

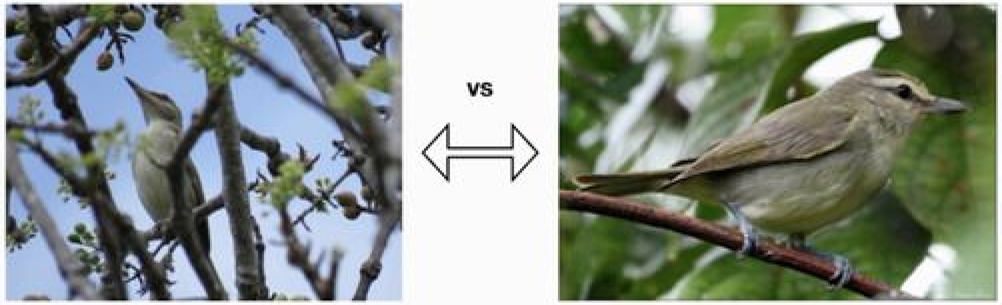
Background

- **Image captioning** [1]



A girl in pink dress is jumping in air.

- **Image comparison** [2]

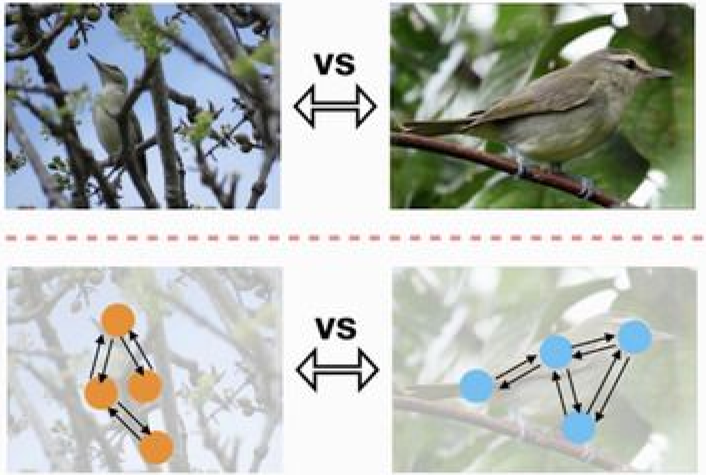


animal1 has a medium sized dark beak,
a white breast and grey wings.
animal2 has a white breast with brown wings and tail,
black eyes and a brown head .

[1] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator.", CVPR 2015
[2] Forbes, Maxwell, et al. "Neural naturalist: generating fine-grained image comparisons.", EMNLP 2019

Motivation

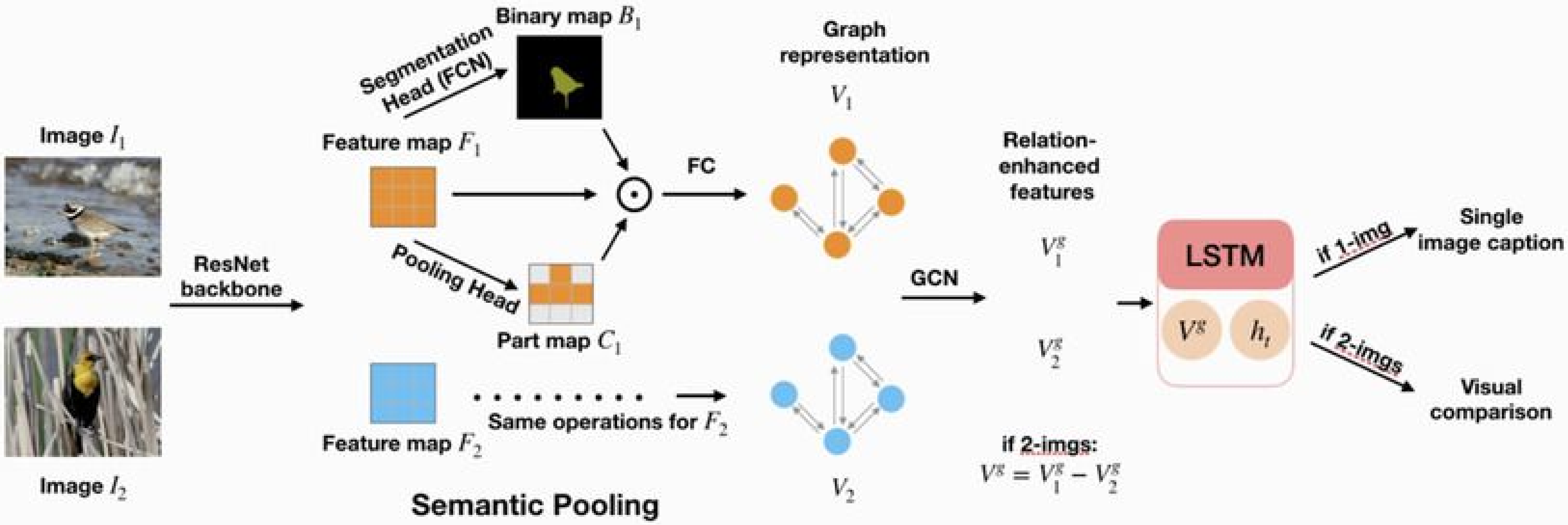
- **Learning semantic representations for each image**



animal1 has a medium sized dark beak, a white breast and grey wings. animal2 has a white breast with brown wings and tail, black eyes and a brown head .

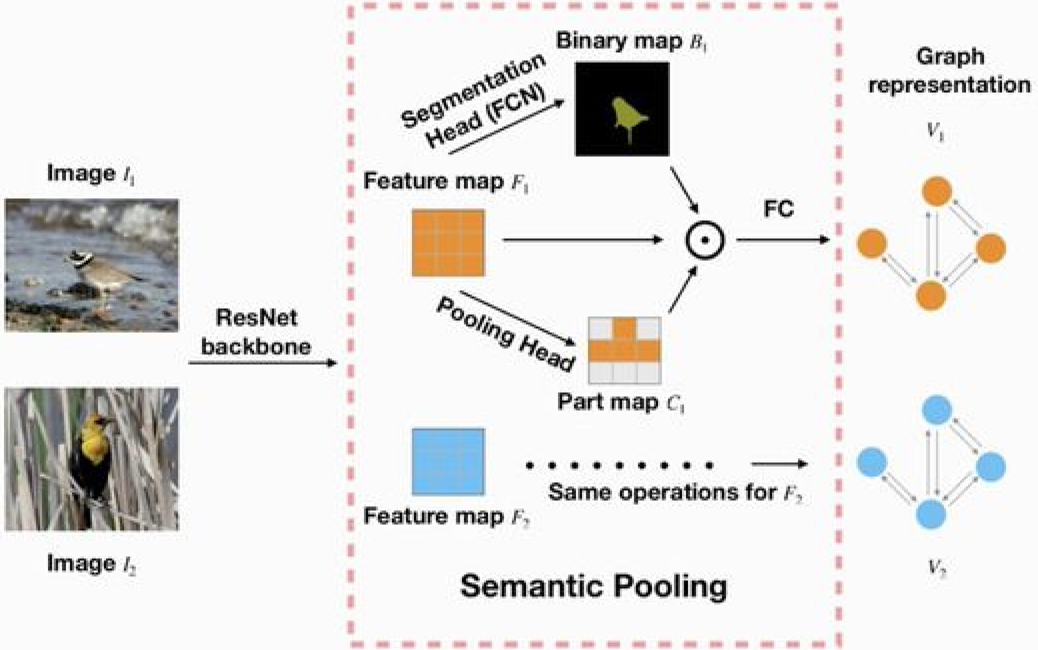
Model

- **Overview**



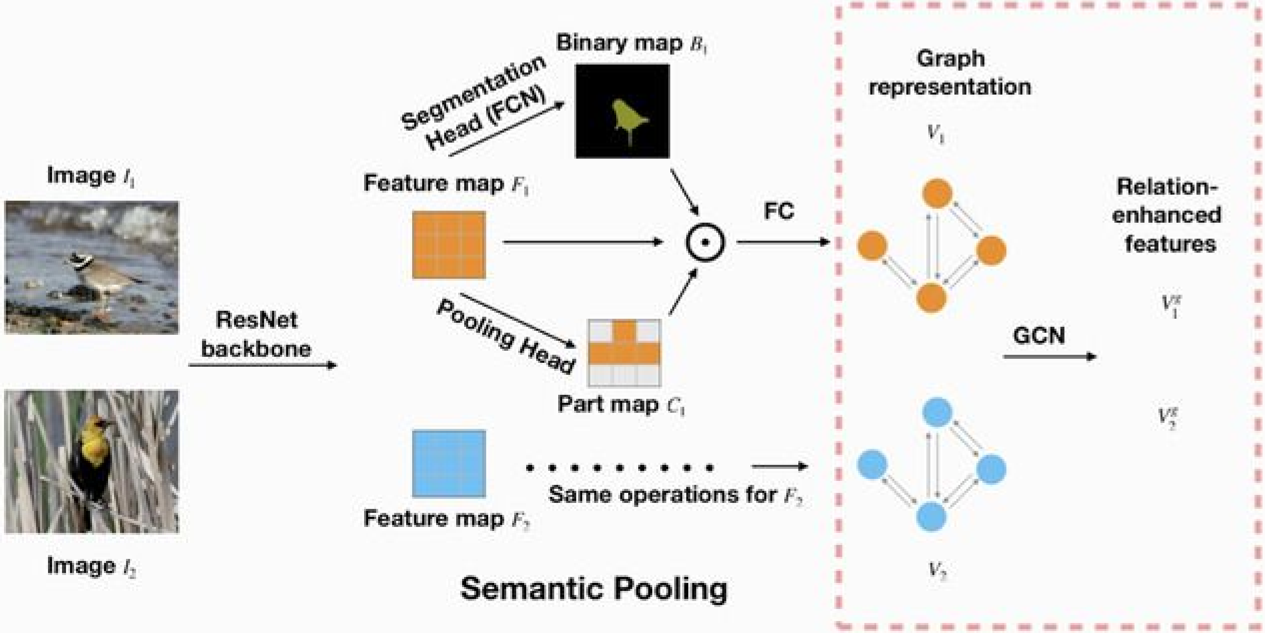
Model

- **Constructing Semantic Representation**



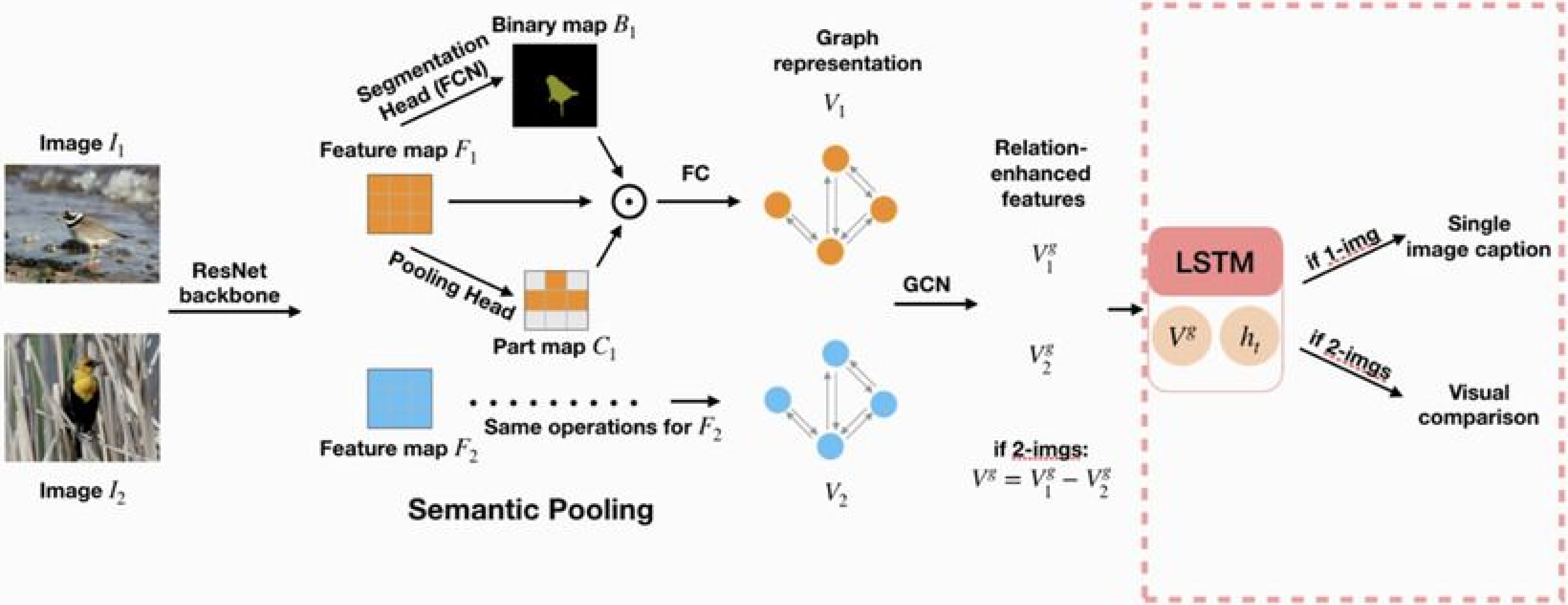
Model

- **Graph Relational Reasoning**



Model

- **Learning to Compare while Learning to Describe**



Experiments

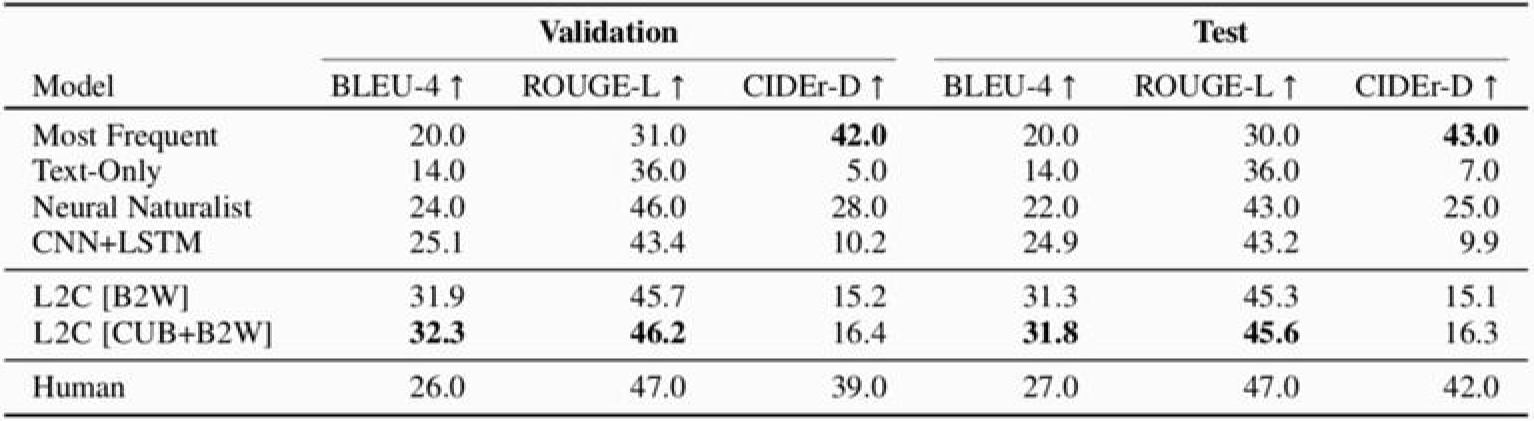
- **Datasets**
  - **Birds-to-Words** [2]
    - **Image comparison, 2 images**
  - **CUB-200-2011** [3]
    - **Image captioning, 1 image**

This bird is mostly black with a bright yellow breast and neck, and orange crown .

[3] Wah, Catherine, et al. "The caltech-ucsd birds-200-2011 dataset." (2011)

Experiments

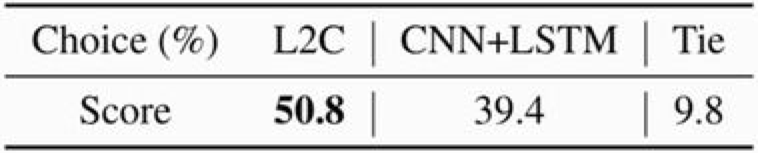
- **Automatic evaluation**



	Validation			Test		
Model	BLEU-4 ↑	ROUGE-L ↑	CIDEr-D ↑	BLEU-4 ↑	ROUGE-L ↑	CIDEr-D ↑
Most Frequent	20.0	31.0	**42.0**	20.0	30.0	**43.0**
Text-Only	14.0	36.0	5.0	14.0	36.0	7.0
Neural Naturalist	24.0	46.0	28.0	22.0	43.0	25.0
CNN+LSTM	25.1	43.4	10.2	24.9	43.2	9.9
L2C [B2W]	31.9	45.7	15.2	31.3	45.3	15.1
L2C [CUB+B2W]	**32.3**	**46.2**	16.4	**31.8**	**45.6**	16.3
Human	26.0	47.0	39.0	27.0	47.0	42.0

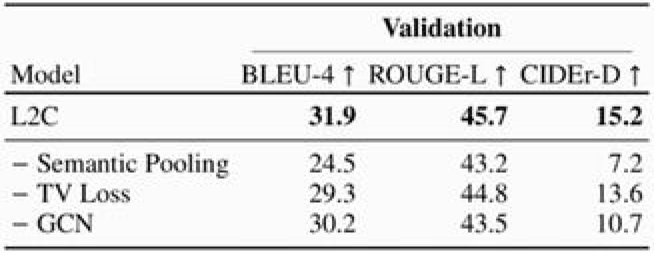
Experiments

- **Human evaluation**
  - Ours vs. CNN+LSTM

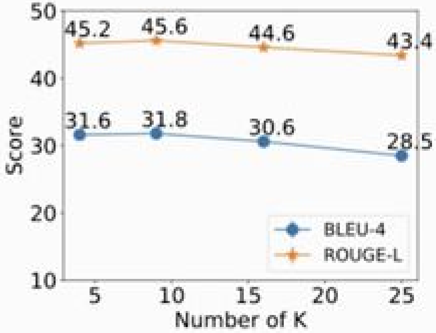


Choice (%)	L2C	CNN+LSTM	Tie
Score	**50.8**	39.4	9.8

- **Ablations**
  - Effect of each module
  - Sensitivity test



	Validation		
Model	BLEU-4 ↑	ROUGE-L ↑	CIDEr-D ↑
L2C	**31.9**	**45.7**	**15.2**
− Semantic Pooling	24.5	43.2	7.2
− TV Loss	29.3	44.8	13.6
− GCN	30.2	43.5	10.7



Conclusions

- This paper presents a learning-to-compare framework for generating visual comparisons .

- Structured image representations can be learned by leveraging segmentation and graph convolutional networks.

- Learning to describe visual differences benefits from understanding the semantics of each image.

Thanks!