Look Twice as Much as You Say: Scene Graph Contrastive Learning for Self-Supervised Image Caption Generation

Chunhui Zhang,[1] Chao Huang,[2] Youhuan Li,[3] Xiangliang Zhang,[4] Yanfang Ye,[4] and Chuxu Zhang[1]

Brandeis University, [1]University of Hong Kong,[2] Hunan University,[3] University of Notre Dame[4]

Motivation & Target

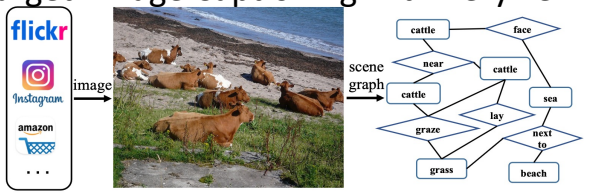
- Motivation: Improving image captioning via leveraging unlabeled images from massive unpaired web sources

- Challenge:

-- 1. Cross-modal data - Unlike previous studies working on single- modal data (e.g., image, text, or graph), image caption generation is a cross-modal task on the intersection of image and text;

-- 2. Complex task - Image caption generation is a complex task that has to generate new content rather than simple classification or prediction task studied in previous work.

- Target: Image Captioning with Very Few Labels



caption generation

GT: **A herd of cattle laying on top of a sandy beach.**

***1% labels* are used:**

VSUA: A cattle a a a a.

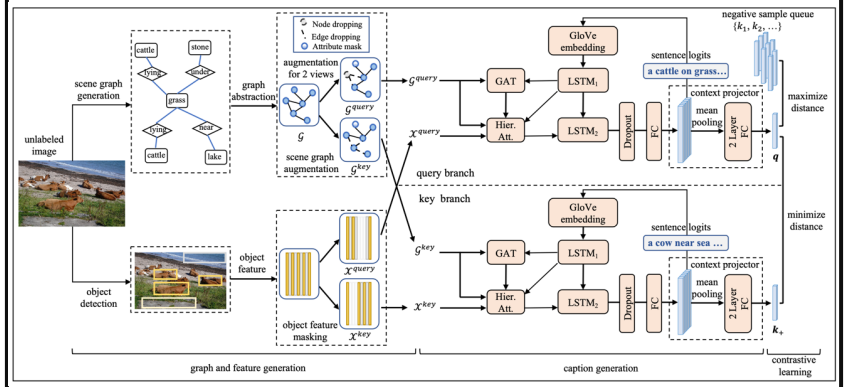
SGAE: A cow is a a a a.

M^2 -T: A herd of sheep standing a a a a a a.

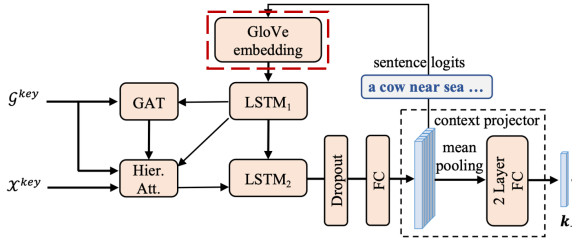
SGCL: A group of cattle grazing in the beach in front of the water.

Model

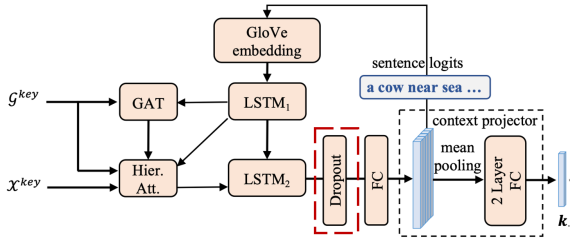
- Overall Framework



-- 1. Pretrained Word Embedding *for NLP Information*



-- 2. Big Dropout Rate *as Semantics Augmentations*



Experiment

- Performance Comparison with SOTAs

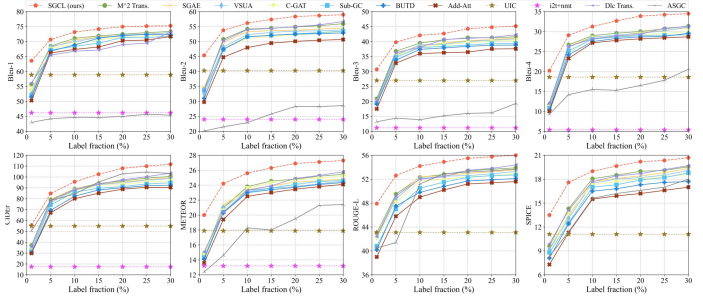


Figure 3: Performances of all models with limited labels (Note that ROUGE-L and SPICE of i2t+nmt are not shown due to missing values in the original work).

- Ablation Study

Table 1: Performances of different model variants with various graph augmentation strategies (Note: N - node dropping, E - edge dropping, A - node attribute masking, O - object feature masking).

Label	N	E	A	O	B-1	B-2	B-3	B-4	C.	M.	R-L	S.
1%				✓	61.8	43.8	28.9	18.2	47.7	18.5	46.3	11.9
	✓			✓	62.5	44.6	30.0	19.1	49.2	19.9	47.0	13.1
		✓		✓	62.5	44.5	29.0	18.5	52.9	19.1	47.3	13.2
			✓	✓	63.1	44.3	28.8	18.6	52.2	19.3	47.2	13.0
	✓	✓	✓	✓	63.0	45.1	29.9	19.3	53.3	19.6	47.5	13.3
	✓	✓	✓	✓	**63.6**	**45.4**	**30.7**	**20.2**	**55.0**	**20.0**	**47.9**	**13.5**
5%				✓	69.4	51.7	36.6	26.2	75.9	22.2	49.4	16.3
	✓			✓	70.3	53.0	38.6	27.9	79.4	23.9	51.9	17.3
		✓		✓	62.5	52.8	38.9	28.5	81.4	19.1	51.9	17.2
			✓	✓	63.1	53.5	38.7	28.6	82.2	19.3	52.0	17.1
	✓	✓	✓	✓	70.3	53.3	39.2	28.1	82.3	24.1	52.2	17.4
	✓	✓	✓	✓	**70.7**	**53.8**	**39.8**	**29.1**	**84.9**	**24.2**	**52.6**	**17.6**

Table 2: Effectiveness of loading pre-trained word embedding (P) and freezing word embedding (F).

Label	P	F	B-1	B-2	B-3	B-4	C.	M.	R-L	S.
1%			62.9	44.8	30.1	19.8	54.1	19.3	47.0	13.0
	✓		63.0	45.1	30.2	19.9	54.4	19.6	47.3	13.0
	✓	✓	**63.6**	**45.4**	**30.7**	**20.2**	**55.0**	**20.0**	**47.9**	**13.5**
5%			69.8	53.2	38.8	28.3	80.0	23.6	52.1	16.9
	✓		70.1	53.6	39.0	28.3	79.8	23.8	52.3	17.1
	✓	✓	**70.7**	**53.8**	**39.8**	**29.1**	**84.9**	**24.2**	**52.6**	**17.6**
10%			71.8	58.3	41.1	33.5	91.7	23.5	52.9	17.5
	✓		72.4	58.3	41.5	33.7	92.6	23.6	53.4	18.3
	✓	✓	**73.2**	**56.2**	**42.1**	**31.3**	**94.6**	**25.6**	**54.2**	**19.0**
20%			74.1	57.8	43.2	33.1	103.9	26.4	54.5	19.6
	✓		74.6	58.0	43.5	33.5	105.5	26.4	55.1	20.0
	✓	✓	**75.0**	**58.4**	**44.3**	**33.9**	**108.0**	**26.9**	**55.5**	**20.2**

- Case Show

***1% labels* are used:**

M^2-T:	A sheep standing in a a a.	A man girl a a a a a.	A girl tennis a a tennis tennis	A elephant of in a a a.	A cow standing standing a a a.
SGAE:	A sheep of standing a a a.	A man is a a a a a.	A man girl a a a a a.	A man is a a a a a.	A sheep of in a a a a a.
VSUA:	A sheep of in a a a.	A man standing a a a a.	A girl girl a a a a a.	A people of a a a a a a.	A cow cow cow a a a a a.
C-GAT:	A group of a a a a.	A man is a a a a a.	A man girl a a a a a.	A street of a a a a.	A sheep of a a a.
SGCL:	**A couple of sheep standing in the grass in a field.**	**A group of people playing a frisbee standing by the sea.**	**A woman is holding a tennis racket on a tennis ball.**	**A group of people standing in a street with a building.**	**A herd of cows walking down a road in the grass.**

***30% labels* are used:**

M^2-T:	A group of sheep standing in a fenced area.	Two men playing frisbee on a dirt field.	A woman is holding a tennis racket in her hand.	A man riding an elephant in front of a building.	A group of cows standing next to each other on a field.
SGAE:	A white sheep is standing in the grass.	A group of men playing a game of frisbee.	A man hitting a tennis ball on a tennis court.	A group of people standing next to an elephant.	A cow standing on top of a lush green field.
VSUA:	A couple of sheep standing next to each other.	A man holding a frisbee in his hand.	Two men playing frisbee on a dirt field.	An elephant standing in front of a building.	A group of cows are standing in the grass.
C-GAT:	A group of sheep grazing in a grassy field.	A man holding a frisbee in his hand.	A woman is playing tennis on the court.	A man riding on the back of an elephant.	A brown cow standing next to a brown cow.
SGCL:	**A couple of sheep standing on a lush green field near a fence.**	**A man is jumping in the air to catch a frisbee on a sea beach.**	**A woman is trying to hitting a tennis ball on a tennis court.**	**An elephant walking down a street with people in the background.**	**A herd of black cows standing next to each other on a lush green field.**




Look Twice as Much as You Say: Scene Graph Contrastive Learning for Self-Supervised Image Caption Generation

Chunhui Zhang[1], Chao Huang[2], Youhuan Li[3], Xiangliang Zhang[4], Yanfang Ye[4], and Chuxu Zhang[1]

Brandeis University[1], University of Hong Kong[2], Hunan University[3], University of Notre Dame[4]