Dual Graph Convolutional Networks with Trasformer and Curriculum Learning for Image Captioning

Xinzhi Dong[1], Chengjiang Long[2*], Wenju Xu[3] , Chunxia Xiao[1*†]

School of Computer Science, Wuhan University[1] , JD Finance America Corporation[2], OPPO US Research Center, InnoPeak Technology Inc.[3]

dongxz97@whu.edu.cn, cjfykx@gmail.com, xuwenju123@gmail.com, cxxiao@whu.edu.cn

ACM multimedia Chengdu, China OCT 20-24 2021

Background & Problem

A stop sign is on a road.

- Image captioning aims to generate a caption from a single image.
- It has excellent development prospects in unmanned driving, unmanned supermarket and other applications.
- Most of the difficulties lie in the accuracy, diversity and complexity of the generated caption.

Motivation

- Transformer has the advantage of parallelization, which is especially critical when generating text.
- Graph neural network (GCN) has excellent performance in extracting the relationship between objects.
- Curriculum learning can make the model train and learn from simple to difficult, and then improve the performance .

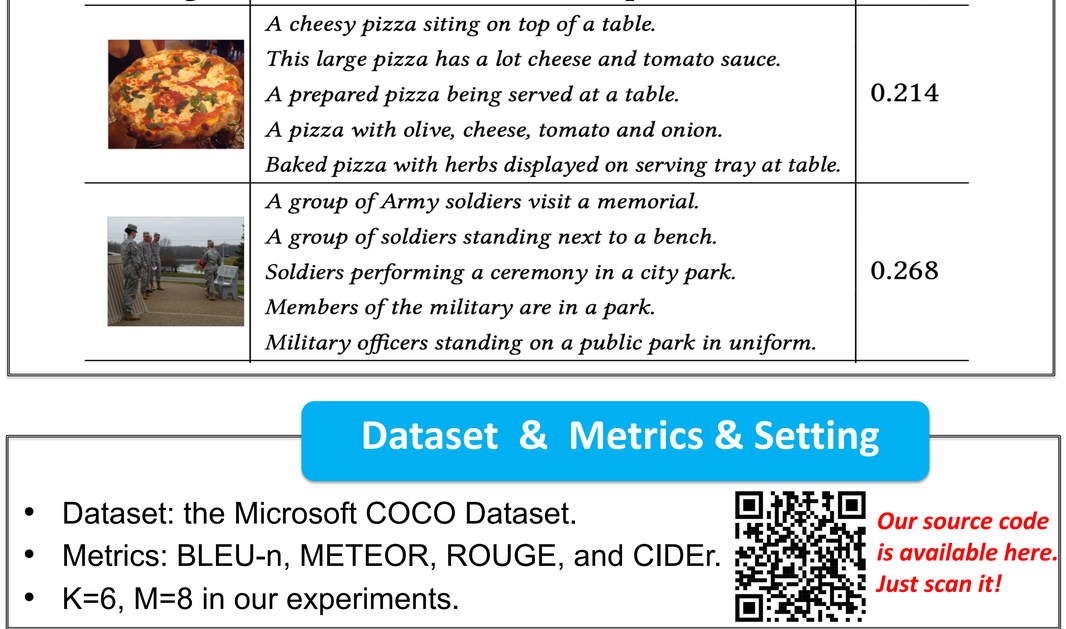
Proposed Method

Dual-GCN+Transformer Framework

Object-Level Encoding | Image-Level Encoding | Transformer as Linguistic Decoder

- Our framework comprises three main parts: Dual-GCN (object-level and image-level) encoder, Transformer decoder, and Curriculum learning.

Curriculum Learning



- We use the cross-review mechanism used to determinethe difficulty level for training examples:

$$DS((I,S)) = \frac{1}{M-1} \sum_{(I,t)\in D_i, k\neq i} \mathcal{E}_k(I,S)$$

Image I	Ground-truth Caption S	$DS(I,S)$
	A cheesy pizza siting on top of a table. *This large pizza has a lot cheese and tomato sauce.* *A prepared pizza being served at a table.* *A pizza with olive, cheese, tomato and onion.* *Baked pizza with herbs displayed on serving tray at table.*	0.214
	A group of Army soldiers visit a memorial. *A group of soldiers standing next to a bench.* *Soldiers performing a ceremony in a city park.* *Members of the military are in a park.* *Military officers standing on a public park in uniform.*	0.268

Dataset & Metrics & Setting

- Dataset: the Microsoft COCO Dataset.
- Metrics: BLEU-n, METEOR, ROUGE, and CIDEr.
- K=6, M=8 in our experiments.

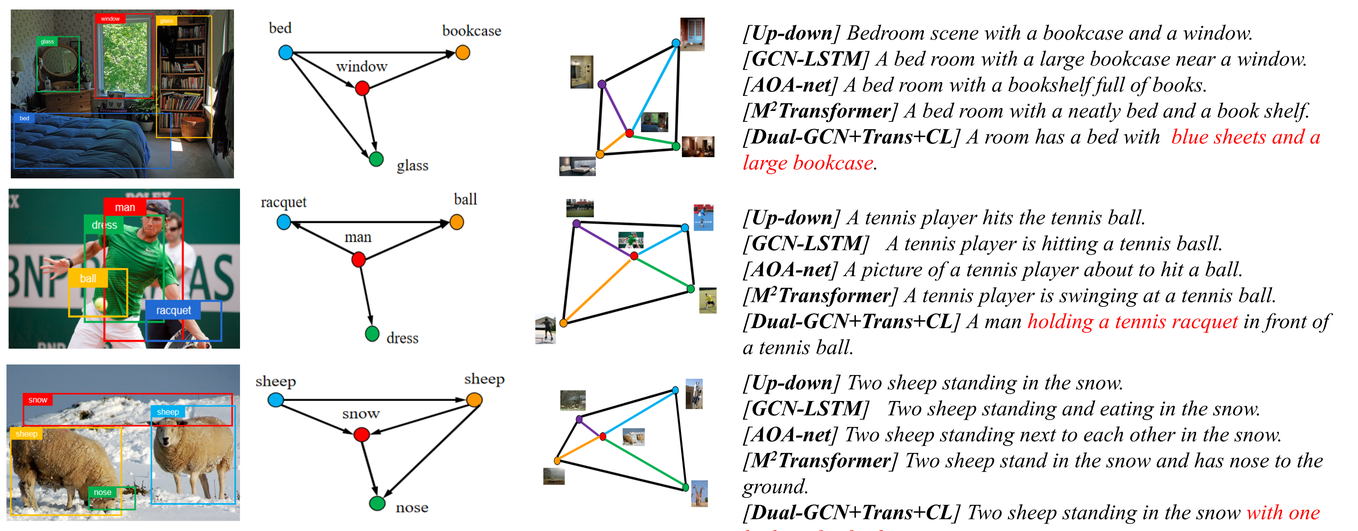
Our source code is available here. Just scan it!

Comparisons with State-of-the-Art

Quantitative Results

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
Up-down [1]	80.2	64.1	49.1	36.9	27.6	57.1	117.9
AOA-NET [21]	81.0	65.8	51.4	39.4	29.1	58.9	126.9
GCN-LSTM [53]	80.8	65.5	50.8	38.7	28.5	58.5	125.3
GCN-LSTM+HIP [54]	81.6	66.2	51.5	39.3	28.8	59.0	127.9
SGAE [50]	81.0	65.6	50.7	38.5	28.2	58.6	123.8
M^2 Transformer [5]	81.6	66.4	51.8	**39.7**	29.4	59.2	127.9
Dual-GCN+Transformer+CL	**82.2**	**67.6**	**52.4**	**39.7**	**29.7**	**59.7**	**129.2**

Visualization Results



*[**Up-down**] Bedroom scene with a bookcase and a window.*
*[**GCN-LSTM**] A bed room with a large bookcase near a window.*
*[**AOA-net**] A bed room with a bookshelf full of books.*
*[**M^2Transformer**] A bed room with a neatly bed and a book shelf.*
*[**Dual-GCN+Trans+CL**] A room has a bed with blue sheets and a large bookcase.*

*[**Up-down**] A tennis player hits the tennis ball.*
*[**GCN-LSTM**] A tennis player is hitting a tennis basll.*
*[**AOA-net**] A picture of a tennis player about to hit a ball.*
*[**M^2Transformer**] A tennis player is swinging at a tennis ball.*
*[**Dual-GCN+Trans+CL**] A man holding a tennis racquet in front of a tennis ball.*

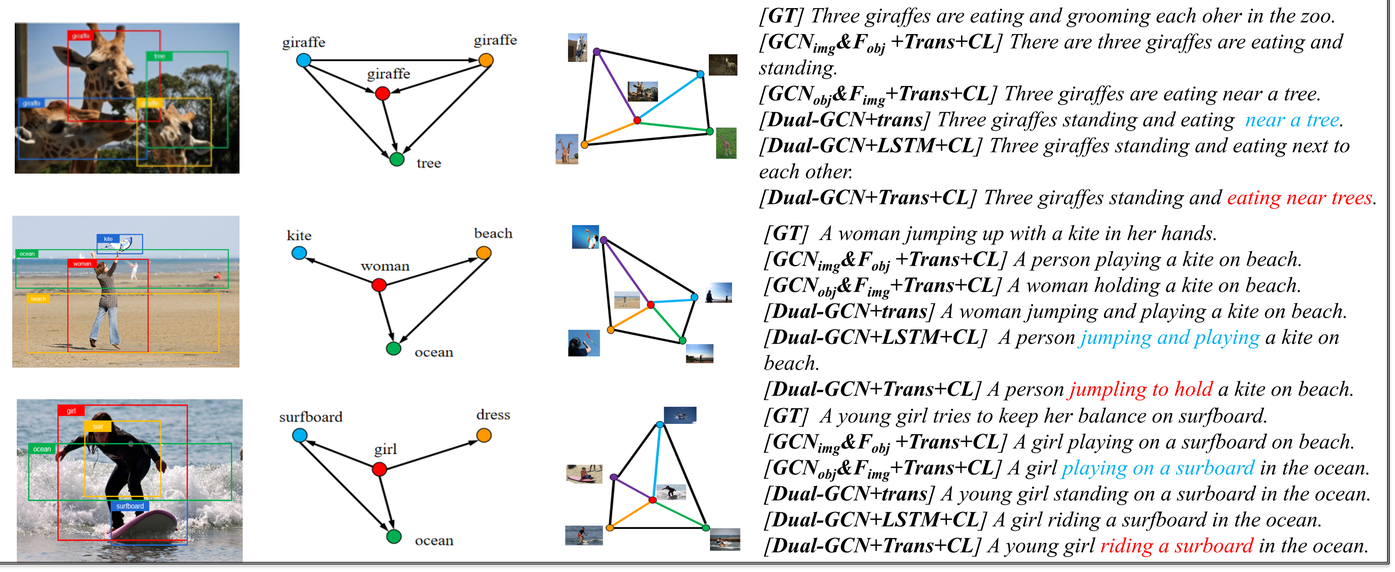
*[**Up-down**] Two sheep standing in the snow.*
*[**GCN-LSTM**] Two sheep standing and eating in the snow.*
*[**AOA-net**] Two sheep standing next to each other in the snow.*
*[**M^2Transformer**] Two sheep stand in the snow and has nose to the ground.*
*[**Dual-GCN+Trans+CL**] Two sheep standing in the snow with one looking for food.*

Abation Study

Quantitative Results

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
F_{obj}+Transformer	71.1	55.9	45.3	28.0	25.2	52.8	114.3
GCN_{obj} + Transformer	80.5	64.7	49.2	37.6	27.9	57.4	125.8
F_{obj}+Transformer+CL	72.3	56.8	46.1	28.4	25.4	53.0	114.7
GCN_{obj} + Transformer +CL	81.6	66.8	51.6	39.3	28.3	58.4	**129.4**
F_{img} + Transformer + CL	80.2	64.1	49.0	37.4	27.4	57.0	124.0
GCN_{img} + Transformer + CL	82.0	66.9	52.2	**39.8**	29.6	59.4	129.0
GCN_{obj}&F_{img} + Transformer + CL	81.9	67.3	52.0	39.5	29.4	59.6	**129.4**
GCN_{img}&F_{obj} + Transformer + CL	82.0	67.3	52.3	39.7	**29.8**	59.6	129.0
Dual-GCN + Transformer	81.8	66.2	51.6	39.2	28.7	58.9	128.0
Dual-GCN + LSTM + CL	81.4	65.8	51.3	38.8	28.0	58.4	127.6
Dual-GCN + Transformer + CL	**82.2**	**67.6**	**52.4**	39.7	29.7	**59.7**	129.2

Visualization Results



*[**GT**] Three giraffes are eating and grooming each oher in the zoo.*
*[**GCN_{img}&F_{obj}+Trans+CL**] There are three giraffes are eating and standing.*
*[**GCN_{obj}&F_{img}+Trans+CL**] Three giraffes are eating near a tree.*
*[**Dual-GCN+trans**] Three giraffes standing and eating near a tree.*
*[**Dual-GCN+LSTM+CL**] Three giraffes standing and eating next to each other.*
*[**Dual-GCN+Trans+CL**] Three giraffes standing and eating near trees.*

*[**GT**] A woman jumping up with a kite in her hands.*
*[**GCN_{img}&F_{obj}+Trans+CL**] A person playing a kite on beach.*
*[**GCN_{obj}&F_{img}+Trans+CL**] A woman holding a kite on beach.*
*[**Dual-GCN+trans**] A woman jumping and playing a kite on beach.*
*[**Dual-GCN+LSTM+CL**] A person jumping and playing a kite on beach.*
*[**Dual-GCN+Trans+CL**] A person jumpling to hold a kite on beach.*

*[**GT**] A young girl tries to keep her balance on surfboard.*
*[**GCN_{img}&F_{obj}+Trans+CL**] A girl playing on a surfboard on beach.*
*[**GCN_{obj}&F_{img}+Trans+CL**] A girl playing on a surboard in the ocean.*
*[**Dual-GCN+trans**] A young girl standing on a surboard in the ocean.*
*[**Dual-GCN+LSTM+CL**] A girl riding a surfboard in the ocean.*
*[**Dual-GCN+Trans+CL**] A young girl riding a surboard in the ocean.*

Failure Cases

Failure cases happen when wrongly selecting images to guide the image captioning of the original image.

*[**GT**] A bunch of red apples and yellow oranges in a closeup.*
*[**Output**] A red apple sitting on top of two apples.*

Conclusion

- A novel framework that takes Dual-GCN as a powerful visual encoder and transformer as the linguistic decoder for image captioning.
- To our best knowledge, we are the first one to apply the Curriculum Learning as the training strategy on image captioning model in an easy-to-hard manner.
- We conduct extensive experiments on the MS COCO dataset. The results strongly demonstrate that our proposed approach outperforms state-of-the-art methods.