Dual Graph Convolutional Networks with Transformer and Curriculum Learning for Image Captioning ——Supplementary Material——

Xinzhi Dong[1], Chengjiang Long[2*], Wenju Xu[3], Chunxia Xiao[1*]
[1] School of Computer, Wuhan University, Wuhan, Hubei, China
[2] JD Finance America Corporation, Mountain View, CA, USA
[3] OPPO US Research Center, InnoPeak Technology Inc, Palo Alto, CA, USA
dongxz97@whu.edu.cn,cjfykx@gmail.com,xuwenju123@gmail.com,cxxiao@whu.edu.cn

ABSTRACT

To demonstrate the effectiveness of our Dual-GCN and Curriculum Learning, and show the limitations of our proposed approach, we provide this supplementary material which mainly contains additional experimental results including **(1)**the more visualization results compared with state-of-the-arts, **(2)** more visualization results of our ablation experiments, **(3)** some visualization results of our similar image selection failure cases and **(4)** some visualization results of our similar image selection success cases. We also provide the running time information. Note that we don't include all these in the central part of the paper due to the space limit.

KEYWORDS

Graph Convolutional Networks, Transformer, Curriculum Learning, Image Captioning

ACM Reference Format:
Xinzhi Dong, Chengjiang Long, Wenju Xu, Chunxia Xiao. 2021. Dual Graph Convolutional Networks with Transformer and Curriculum Learning for Image Captioning ——Supplementary Material——. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Chengdu, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 ADDITIONAL EXPERIMENTAL RESULTS

1.1 More visualization results compared with state-of-the-arts

The more visualization results are provided in Figure 1. We compare our proposed method *Dual-GCN+Transformer+CL* with Updown [1], AOA-NET [3], GCN-LSTM [4] and M^2 Transformer [2]. Obviously, our proposed method *Dual-GCN+Transformer+CL* obtains better performance in image captioning.

For the sake of brevity, we temporarily choose the number of nodes as **4** and **5** in the object-level graph and image-level graph, respectively. Note that there might be many nodes selected on the object level, but no more than 36 at most. Similarly, there can be many similar images in the image level selection, but in this experiment, we set it to 8. Figure 1 shows that our *Dual-GCN+Transformer+CL* model can generate more descriptive and accurate sentences, including the supplementary description of visual objects and the generation of non-visual words. This is because GCN_{obj} encodes more detailed information between different objects in an original image, making some objects their interaction behavior not lost.

Similarly, GCN_{img} encodes the information between multiple similar images and uses the global features to guide the text generation. This is because images with high similarity tend to have the same subject-object and background area. This information can enhance the model's understanding of the visual features and strengthen the integration of visual features and text features for text generation.

1.2 More Visualization results of ablation experiments

In our framework, there are four changeable factors. *i.e.*, *Transformer*, *Dual-GCN* including GCN_{obj} and GCN_{img}, and *Curriculum Learning*. Therefore, strictly based on the control variable method, we conduct several ablation experiences. Figure 2 shows more visualization results. From the figure, we can see that our model *Dual-GCN+Transformer+CL* can generate more detailed sentences. For example, in the fourth image, *Dual-GCN+Transformer+CL* generates **"sitting at a counter and talking"**, while the other variants cannot. With any factor changes, either GCN_{img} or *Curriculum Learning*, it may lead to a less reasonable caption generated. All these observations have clearly verified the validity of *Dual-GCN*, *Transformer*, and *Curriculum Learning*.

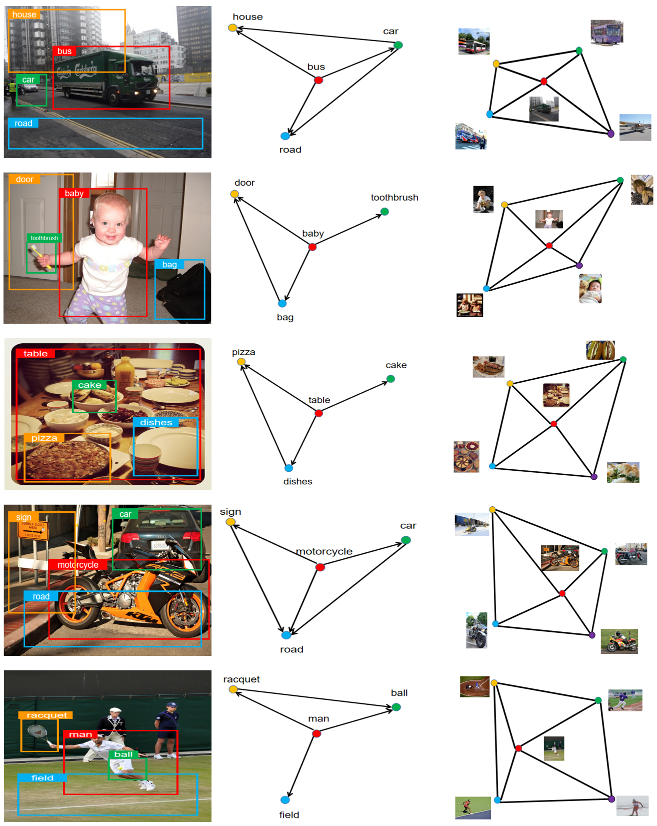
1.3 More visualization results of high-quality top-K images

To experimentally improve better text generation when the selected multiple similar images are more accurate, we present some visualization results with high-quality contextual images as the top-K ($K = 8$) image as success cases, as shown in Figure 3.

*This work was co-supervised by Chengjiang Long and Chunxia Xiao. Chunxia Xiao is the corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10.
https://doi.org/10.1145/1122445.1122456



[Up-down] A large long bus on a city road.
[AOA-net] A bus driving on the street with a car.
[GCN-LSTM] A white and green bus driving on a road.
[M^2 Transformer] A green bus driving near a car.
[Dual-GCN+Trans+CL] A large green bus driving on the road in front of a car.

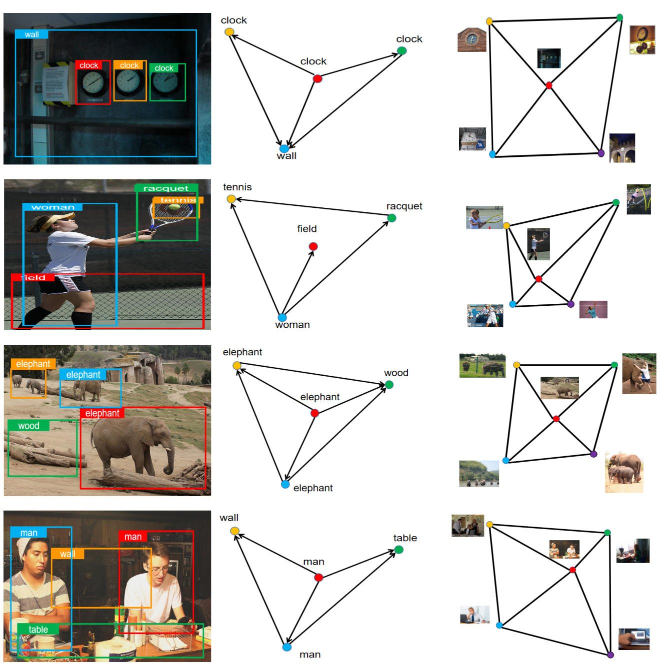
[Up-down] A baby standing with a toothbrush.
[AOA-net] A baby running with a toothbrush near a bag.
[GCN-LSTM] A baby is look forward with a toothbrush.
[M^2 Transformer] A baby is holding a toothbrush.
[Dual-GCN+Trans+CL] A baby simling and holding a toothbrush in his hand .

[Up-down] A table with a pizza and some juice.
[AOA-net] A table full of pizza and bread.
[GCN-LSTM] A table with a pizza and somg bread on it.
[M^2 Transformer] There are some pizza and cakes on the table.
[Dual-GCN+Trans+CL] A large wooden table with some pizza and deserts on it.

[Up-down] A motorcycle parked on the road.
[AOA-net] A motorcycle parked near a silver car.
[GCN-LSTM] A motorcycle parked near a car.
[M^2 Transformer] A motorcycle parked in front of a car on the road.
[Dual-GCN+Trans+CL] A motorcycle parked near a car and a sign on the road.

[Up-down] A man hitting a tennis ball.
[AOA-net] A man playing a tennis with a racquet.
[GCN-LSTM] A player is hitting a tennis ball in the field.
[M^2 Transformer] A player hitting a tennis ball in a game.
[Dual-GCN+Trans+CL] A player using a racquet to hit a tennis ball in the field.

Figure 1: Visualization results compared with four state-of-the-arts. Note that we only plot top-4 images for simplicity.



[GT] Three clocks on a wall, all with different times.
[GCN_{img}&F_{obj}+Trans+CL] A large wall with three clocks on it.
[GCN_{obj}&F_{img}+Trans+CL] Three clocks are hung on the wall.
[Dual-GCN+Trans] There are three clocks on the wall.
[Dual-GCN+LSTM+CL] There are three different clocks on the wall.
[Dual-GCN+Trans+CL] There are three different clocks on a wall in the room.

[GT] A woman on a tennis court holding a tennis racket.
[GCN_{img}&F_{obj}+Trans+CL] A player hitting a tennis ball in the field.
[GCN_{obj}&F_{img}+Trans+CL] A woman playing tennis ball in a game.
[Dual-GCN+Trans] A woman hitting a tennis ball with a racquet.
[Dual-GCN+LSTM+CL] A woman hitting a tennis ball with her racket.
[Dual-GCN+Trans+CL] A woman swinging a racket to hit a tennis ball in a game.

[GT] A group of elephants walking across a grass covered field.
[GCN_{img}&F_{obj}+Trans+CL] Several elephants standing near a wood.
[GCN_{obj}&F_{img}+Trans+CL] There are three elephants standing next to each other.
[Dual-GCN+Trans] An elephant standing beside an elephant.
[Dual-GCN+LSTM+CL] Elephants stand in the field next to each other.
[Dual-GCN+Trans+CL] There are group of elephants eating next to a wood in the field.

[GT] Two young men sit at a kitchen bar covered with many things.
[GCN_{img}&F_{obj}+Trans+CL] Two young men sitting at a table next to each other.
[GCN_{obj}&F_{img}+Trans+CL] Two men standing at a bar together.
[Dual-GCN+Trans]There are two men talking and sitting next to each other.
[Dual-GCN+LSTM+CL]Two men sitting near each other at a table.
[Dual-GCN+Trans+CL]Two men sitting at a counter and talking .

Figure 2: Visualization results of ablation experiments. Note that we only plot top-4 images for simplicity.

Input Image

[GT] *A bowl filled with lots of oranges on a counter .*
[Output] *There is a bowl filled with some oranges.*

Input Image

[GT] *A man riding skis down a snow covered ski slope.*
[Output] *A man is skiing and coming down in front of the slope.*

Figure 3: The result of similar image selection success.

1.4 More visualization results of low-quality top-K images

Although images with high similarity can provide certain visual information, distinguishing similar images accurately remains a complex problem. In our proposed method, there might be errors in the selection of a similar image sometimes. Figure 4 shows some visual examples of selection failure. We can find that sometimes the model selects visually irrelevant images for guidance. In this case, it leads to inaccuracy and even some errors in text generation. Apparently, this suggests that there still exist space and potential for our proposed method to improve.

2 ADDITIONAL ANALYSIS

2.1 Running time

At the testing stage, we take our model to generate captions of 100 images. It totally took about one minute. The average time to generate caption for a single image is 0.6 second.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 6077–6086.

[2] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE International Computer Vision and Pattern Recognition (CVPR)*. 10578–10587.

[3] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on Attention for Image Captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 4633–4642.

[4] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring Visual Relationship for Image Captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Vol. 11218. 711–727.

[GT] *A man riding a surfboard in the ocean.*
[Output] *A man riding a surfboard on a boat.*

[GT] *A public bus parked at a bus stop .*
[Output] *A red bus sitting next to a train.*

Figure 4: The result of similar image selection failure.