

A-CAP: Anticipation Captioning with Commonsense Knowledge

Supplemental Material

I. Prompt Learning

The prompt learning was developed by NLP research [41, 42, 43]. It considers pre-trained language models such as BERT [10], as knowledge-based sources of useful information for downstream tasks. The key idea is to create a prompt (template) that can guide the pre-trained model through the adaptation process to a new task. It should be noted that the prompt format should be the same as the input format learned by the pre-trained model. Furthermore, the parameters of the pre-trained model are not updated during the training process; instead, we train the layers to learn prompt embeddings. The concept of prompt learning has recently been explored in computer vision [39, 40], where the context-word-generated prompt is converted into a set of learnable vectors and fed into a pre-trained vision-language model to solve downstream tasks.

In our method, we use prompt learning in the same way as recent methods [39, 40]. We see that the key idea of VinVL [38] is the usage of concepts (object names), which allows better alignment between vision and language spaces, leading to the appearance of concepts in the caption. If we add forecasted concepts to the model, the model will be able to generate the caption based on the forecasted concepts. In our method, we combine detected and forecasted concepts to create the prompt. To this end, we change the VinVL’s input to words–(detected, forecasted)concepts–ROIs because the format of the prompt should be familiar to the pre-trained model (i.e., sequence of words–concepts–ROIs). During the training time, by using cross-entropy loss, we update the graph neural network to learn the embeddings for the concepts to ensure that the pre-trained model can understand the prompt embeddings. After training, the pre-trained model can easily generate the desired captions from the input.

II. More Examples

We randomly select more examples of captions generated by our method and our compared methods. They are shown in Figs. A, C, E, and G. We also show their corresponding generated images obtained by using stable diffusion model [28] in Figs. B, D, F, and H. Along with Fig. 3 in

the main paper, these figures consistently demonstrate that our method generates captions that are more accurate, descriptive, and plausible than the other methods.

In addition, Figs. I and J show the captions generated by ablated models: A-CAP w/o GNN, A-CAP w/o context, and our full model. We can see that, as stated in the main paper, the captions generated by A-CAP w/o GNN most likely describe the inputs, whereas those generated by A-CAP w/o context are far from the inputs. Meanwhile, our full model can produce plausible captions.

The observations from the additional examples support our conclusion that our method is better suited to the anticipation captioning task than the other methods and ablated models.

III. Visualization of Knowledge Graph

We visualize the knowledge graphs corresponding to the examples in Fig. 3 (main paper) in Figs. K, L, M, and N to better understand the contributions of forecasted concepts in the anticipated captions. The left graph in each figure is the full knowledge graph, which contains all detected and forecasted concepts. We see nodes in the graph are densely connected, meaning most nodes are related. We remark that the number of nodes is 100 ($= 4 \times 10 + 60$) and the number of edges is 6000 on average.

The right graph, on the other hand, is the portion of the knowledge graph that is extracted using only the forecasted concepts (brown nodes) appearing in the anticipated caption and the detected concepts (blue nodes) related to the forecasted ones. We can see that our method successfully retrieves forecasted concepts from ConceptNet [30], which are the future of detected concepts. More importantly, our method can include forecasted concepts in the final caption thanks to our usage of prompt learning.

References

- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019.
- [17] Mike Lewis, Yinhan Liu, Naman Goyal, Mar-

- jan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In ACL, 2020.
- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022.
- [30] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In AAAI, 2017.
- [35] Xin Wang, Wenhua Chen, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In ACL, 2018.
- [38] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. In CVPR, 2021.
- [39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In CVPR, 2022.
- [40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. IJCV, 2022.
- [41] Gao Tianyu, Fisch Adam, and Chen Danqi. Making Pre-trained Language Models Better Few-shot Learners. ACL, 2021.
- [42] Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. EMNLP, 2021.
- [43] Xiang Lisa Li, and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. ACL, 2021.









Sparsely temporally-ordered images	Oracle image (for reference)	VinVL	VinVL + Oracle image	AREL + BART	A-CAP	Ground-truth
		everyone had a great time.	it was a great night with lots of great food.	they were all excited for the event to start.	[female] was excited to see what was going on.	this is [female] watching as her dad scores a strike to win the game.
		the couple was wed and the day was merry.	the bride's parents were especially happy.	at the end of the night, they all took a group photo to remember the day.	the bride and groom made a toast to their new life together.	the pastor presenting them as husband and wife for the first time.
		they finished the night with a glass of wine and some good food.	the lights were still on by the time we were done.	we had a great time and enjoyed it.	after a long day of drinking, they went home for the night.	later they all left, full and happy
		the band played with purple lights, stunning the audience.	at the end of the day, we all went home.	the audience was having a great time.	the stage was set up and everyone was ready for the show to begin.	Afterwards the graduates congratulated each other and discussed the directions their lives were now going in beyond school.

Figure A. Examples of generated captions obtained by all compared methods. We show the oracle images and ground-truth captions for reference purposes. VinVL [38] generates captions that are out of context with the input images. VinVL [38] + Oracle image sometimes generates reasonable captions. AREL [35] + BART [17] tends to generate a general ending for the sequence of images. On the other hand, our method A-CAP predicts more accurate, descriptive, and plausible captions than others.





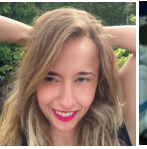





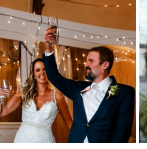


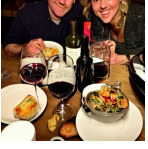
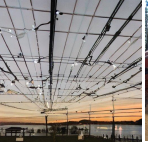




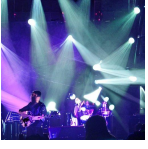


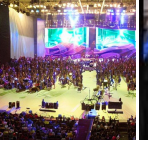

Sparsely temporally-ordered images	VinVL	VinVL + Oracle image	AREL + BART	A-CAP	Ground-truth
					
					
					
					

Figure B. The generated images obtained by using stable diffusion model [28] to generate an image from each generated caption in Fig. A. The order of images is the same as the order of captions in Fig. A. The images generated using our captions are close to the ground-truth ones while those by other methods are not.

Sparsely temporally-ordered images	Oracle image (for reference)	VinVL	VinVL + Oracle image	AREL + BART	A-CAP	Ground-truth
		and, to make it less boring, goof around with some of the products.	and of course, there was no food left for the people to eat.	he wasn't sure if he wanted to buy it, but he didn't want to buy it at the store.	finally, we got to grab some lunch at the convenience store.	time to check out the food.
		they were there to protest against laws they didn't accept	and they paid their respects to those who had lost their lives.	i had a great time.	the government officials also honored the fallen soldiers.	It was easy to tell that he was impressed about the special event for veterans' day
		this is the stuff I got to keep in the museum.	the jack - o - lantern competition had some very impressive contestants.	they have a great time and are ready to go home.	at the end of the night, they all made their own pumpkins.	while others made their own, unique creations.
		the stands selling sculptures are also very popular.	we went to the waffle shop to get some food.	i really enjoyed all of the fresh fruits and vegetables.	we got to buy some fresh fruit and vegetables.	after lunch, we found a small candy shop and got some dessert.

Figure C. Examples of generated captions obtained by all compared methods. We show the oracle images and ground-truth captions for reference purposes. VinVL [38] generates captions that are out of context with the input images. VinVL [38] + Oracle image sometimes generates reasonable captions. AREL [35] + BART [17] tends to generate a general ending for the sequence of images. On the other hand, our method A-CAP predicts more accurate, descriptive, and plausible captions than others.

Sparsely temporally-ordered images	VinVL	VinVL + Oracle image	AREL + BART	A-CAP	Ground-truth

Figure D. The generated images obtained by using stable diffusion model [28] to generate an image from each generated caption in Fig. C. The order of images is the same as the order of captions in Fig. C. The images generated using our captions are close to the ground-truth ones while those by other methods are not.

Sparsely temporally-ordered images	Oracle image (for reference)	VinVL	VinVL + Oracle image	AREL + BART	A-CAP	Ground-truth
		the celebration began to wind down as the night dragged on.	we had a great time.	the men are waiting for the food to be served.	the night ended with people camping in tents on the side of the road.	everyone made friends.
		we had a great time watching the game.	the owner of the car was awarded a star.	afterward there were a lot of rides.	the arena was packed with people.	self driving cars are shown to be the future of transportation.
		after the ceremony, the teams got to eat outside.	and got to throw the ball.	the game was very exciting.	we had a great time at the baseball game.	he was tired by the end of the day, but it is a day he will remember forever.
		the night ended with an amazing fireworks display.	the weather was nice and we had a great time.	they had a lot of fun.	we had a great time at the concert.	when the show starts everybody is cooking, eating and waiting for the speaker to begin.

Figure E. Examples of generated captions obtained by all compared methods. We show the oracle images and ground-truth captions for reference purposes. VinVL [38] generates captions that are out of context with the input images. VinVL [38] + Oracle image sometimes generates reasonable captions. AREL [35] + BART [17] tends to generate a general ending for the sequence of images. On the other hand, our method A-CAP predicts more accurate, descriptive, and plausible captions than others.








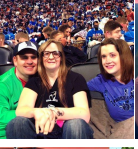

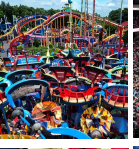
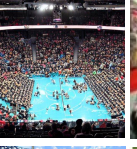



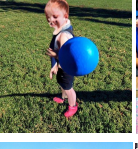
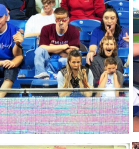
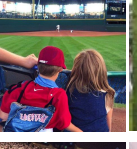





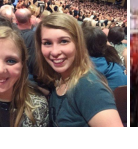

Sparsely temporally-ordered images	VinVL	VinVL + Oracle image	AREL + BART	A-CAP	Ground-truth
					
					
					
					

Figure F. The generated images obtained by using stable diffusion model [28] to generate an image from each generated caption in Fig. E. The order of images is the same as the order of captions in Fig. E. The images generated using our captions are close to the ground-truth ones while those by other methods are not.

Sparsely temporally-ordered images	Oracle image (for reference)	VinVL	VinVL + Oracle image	AREL + BART	A-CAP	Ground-truth
		the park was beginning to close because it was getting late, so they left.	we had a great time and can't wait to come back next year.	they won a prize at a prize.	it was a great night and everyone had a good time.	the lobsters are then place in the tubs for the guest to choose from.
		my husband loves a good organization.	the principal gave a presentation for the students.	they had a great time at the wedding.	[male] is tired from all the school day and is ready to go home.	he studied throughout the day, making sure to finish his homework.
		this is the stuff i got to keep in the museum.	everyone had a great time.	after that we played in the water.	the day ended with a picnic in the park.	there were also games for children.
		the view from our room is breathtaking.	the boy was really glad he got out of the car.	he rode his bike down the road and stopped for a bite to eat.	as the sun went down, we headed back home.	I think I'll take a nap and ride my bike some more tomorrow.

Figure G. Examples of generated captions obtained by all compared methods. We show the oracle images and ground-truth captions for reference purposes. VinVL [38] generates captions that are out of context with the input images. VinVL [38] + Oracle image sometimes generates reasonable captions. AREL [35] + BART [17] tends to generate a general ending for the sequence of images. On the other hand, our method A-CAP predicts more accurate, descriptive, and plausible captions than others.

Sparsely temporally-ordered images	VinVL	VinVL + Oracle image	AREL + BART	A-CAP	Ground-truth

Figure H. The generated images obtained by using stable diffusion model [28] to generate an image from each generated caption in Fig. G. The order of images is the same as the order of captions in Fig. G. The images generated using our captions are close to the ground-truth ones while those by other methods are not.



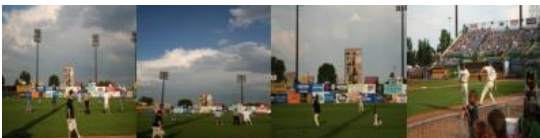





Sparsely temporally-ordered images	Oracle image (for reference)	A-CAP w/o GNN	A-CAP w/o context	A-CAP (full)	Ground-truth
		everyone had a good time carving pumpkins.	they made pumpkins for everyone to eat.	at the end of the night, we got to see some scary pumpkins.	what a harvest at halloween.
		it was a great day for both teams.	we had a great time.	after the game, the mascots posed for a picture with their fans.	the first inning was over and the pitcher was practicing.
		they had a lot of fun in the car, and they had a lot of fun.	the driver of the yellow car happily takes a picture with the camera crew to celebrate his win	the driver of the red and yellow car took the lead in a race.	it is the final lap and the lowe's car is in the lead, the crowd goes wide.
		there were many people walking around in the parade.	we had a great time.	the parade ended in the center of the city.	all in all, it was a fun celebration .

Figure I. Examples of generated captions by two ablated models: A-CAP w/o GNN, A-CAP w/o context, and full model A-CAP. We select two inputs where the detected concepts almost overlap. A-CAP w/o GNN generates captions that most likely describe the inputs. A-CAP w/o context generates captions that are far from the inputs and similar to each other.









Sparsely temporally-ordered images	Oracle image (for reference)	A-CAP w/o GNN	A-CAP w/o context	A-CAP (full)	Ground-truth
		i had a great time there.	the family of the bride and groom were happy to be together.	the family and friends gathered for a group photo.	however in the end seeing his happy face brought a smile to mine
		the family was incredibly proud of her achievement.	the family and friends pose for one last photo before heading off to the hospital.	after the ceremony, the family and friends posed for a picture on the steps.	many friends and family have come to show their support.
		the food was great and i had a great time.	the fruit and vegetables were very colorful.	we ended the night with a ride on the horse and buggy.	our evening ended with a carriage ride back to our hotel.
		the downtown streets were lined with people enjoying themselves	we ended the day with a ride on a boat on the river.	then we headed back to the market to buy some fresh fruit.	they all looked so good and fresh.

Figure J. Examples of generated captions by two ablated models: A-CAP w/o GNN, A-CAP w/o context, and full model A-CAP. We select two inputs where the detected concepts almost overlap. A-CAP w/o GNN generates captions that most likely describe the inputs. A-CAP w/o context generates captions that are far from the inputs and similar to each other.

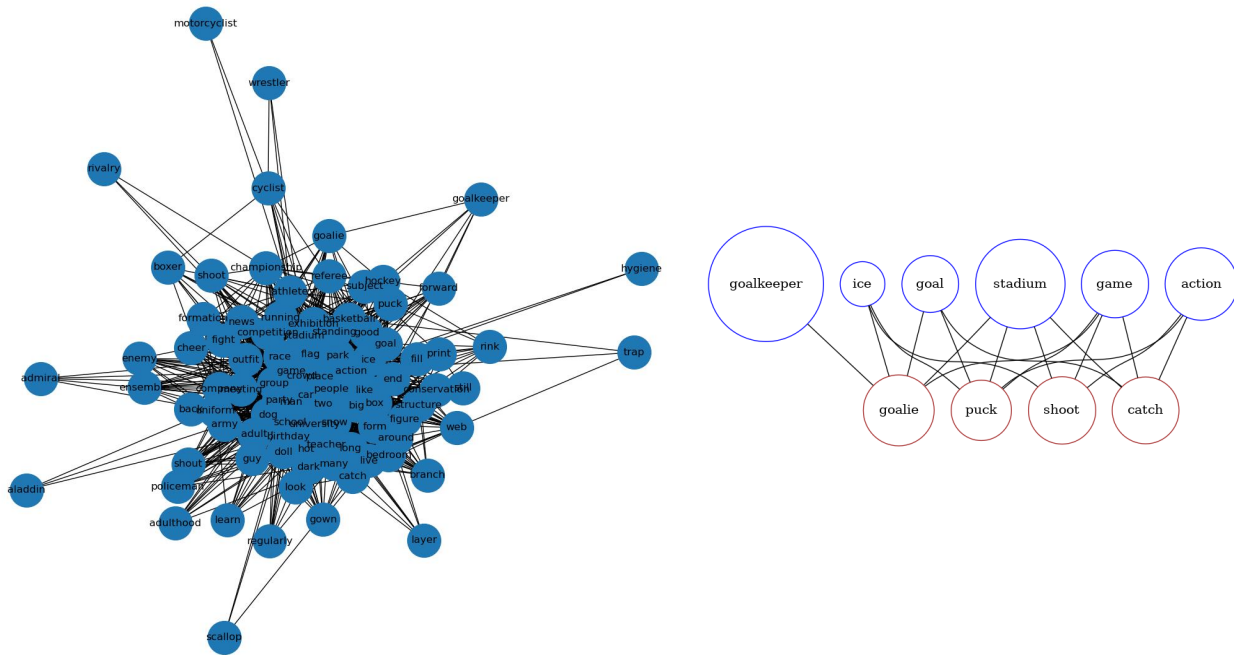


Figure K. Visualization of knowledge graph of the first example in Fig.3 in the main paper. The full graph is shown on the left, while the detected concepts (blue nodes) and forecasted concepts (brown nodes) that contribute to the caption are shown on the right. We can see that our method successfully retrieves forecasted concepts from ConceptNet [30], which are the future of detected concepts.

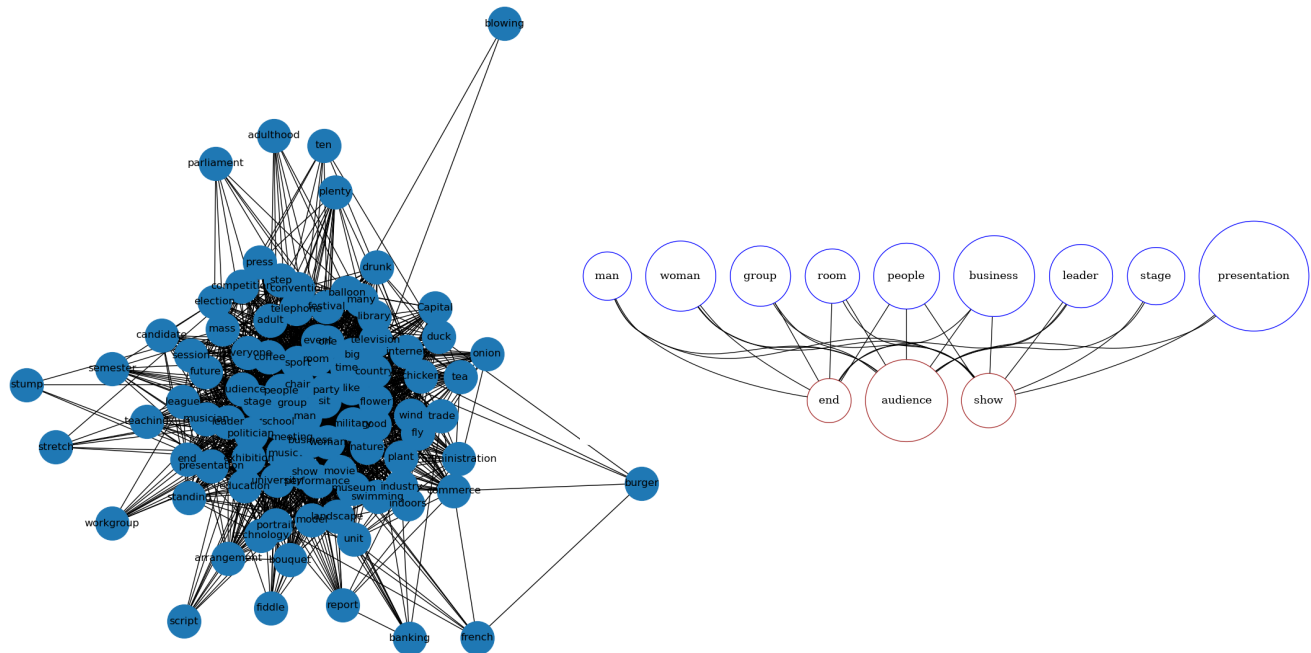


Figure L. Visualization of knowledge graph of the second example in Fig.3 in the main paper. The full graph is shown on the left, while the detected concepts (blue nodes) and forecasted concepts (brown nodes) that contribute to the caption are shown on the right. We can see that our method successfully retrieves forecasted concepts from ConceptNet [30], which are the future of detected concepts.

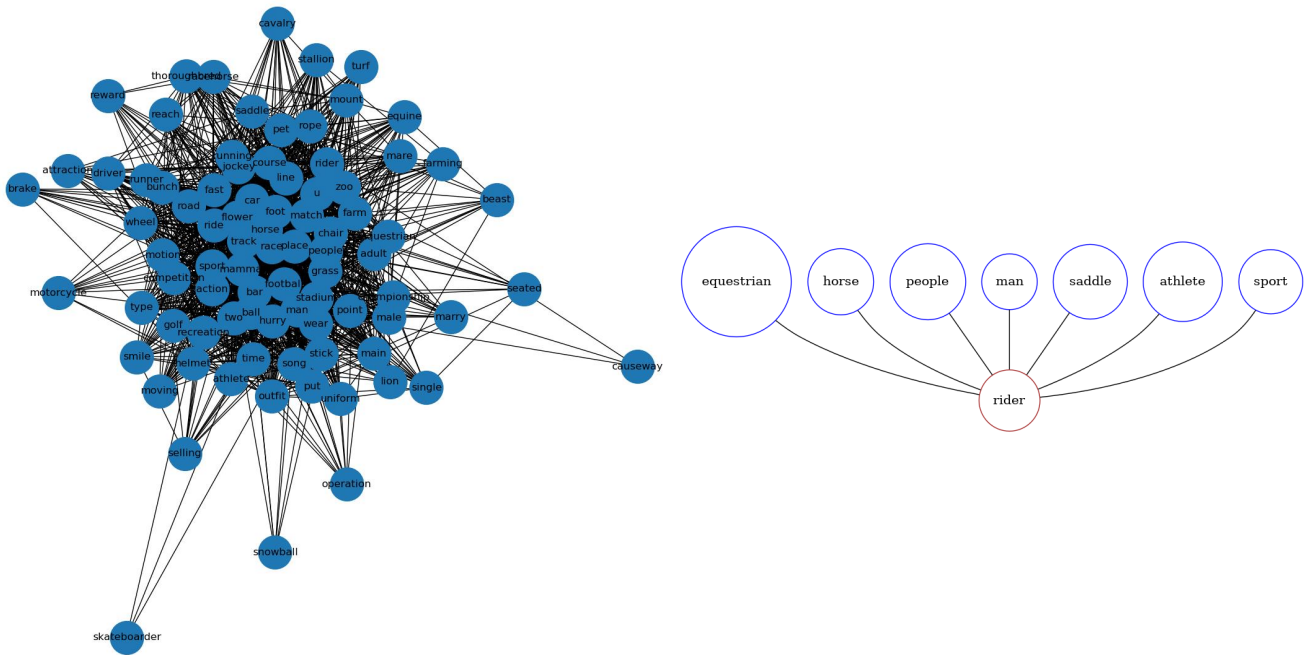


Figure M. Visualization of knowledge graph of the third example in Fig.3 in the main paper. The full graph is shown on the left, while the detected concepts (blue nodes) and forecasted concepts (brown nodes) that contribute to the caption are shown on the right. We can see that our method successfully retrieves forecasted concepts from ConceptNet [30], which are the future of detected concepts.

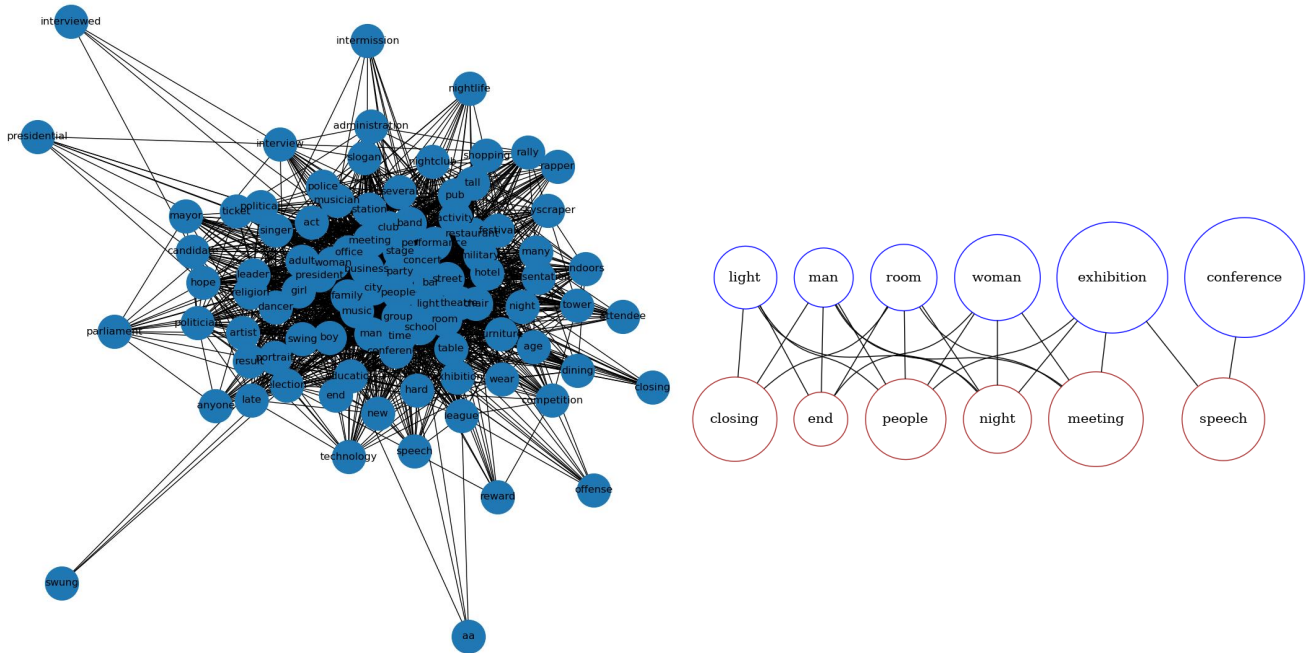


Figure N. Visualization of knowledge graph of the fourth example in Fig.3 in the main paper. The full graph is shown on the left, while the detected concepts (blue nodes) and forecasted concepts (brown nodes) that contribute to the caption are shown on the right. We can see that our method successfully retrieves forecasted concepts from ConceptNet [30], which are the future of detected concepts.