KM-BART: Knowledge Enhanced Multimodal BART for Visual Commonsense Generation

Yiran Xing*♠ Zai Shi*♣ Zhao Meng*♣
Gerhard Lakemeyer♠ Yunpu Ma♦ Roger Wattenhofer♣
♠RWTH Aachen, Germany
♦LMU Munich, Germany
♣ETH Zurich, Switzerland
{yiran.xing, gerhard}@rwth-aachen.de
cognitive.yunpu@gmail.com
{zaishi, zhmeng, wattenhofer}@ethz.ch

1 Implementation Details

Our KM-BART is based on HuggingFace Transformers[1] and PyTorch[2]. For all our experiments, we use PyTorch built-in automatic mixed precision to speed up training. Our model has about 141 million parameters.

1.1 Pretraining

In pretraining, we use the AdamW optimizer with a learning rate of 1e-5. We use a dropout rate of 0.1 for regularization in fully connected layers. We pretrain our model for 20 epochs under each of the pretraining settings. We conduct our pretraining on 4 Titan RTX GPUs with an effective batch size of 256. Pretraining the model with all four pretraining tasks takes around one week. For our full model, we set the loss weights $W_{KCG}, W_{AP}, W_{RP}, W_{MRM}$ to 1.0 and W_{MLM} to 5.0.

1.2 Finetuning

We use the same optimizer and learning rate during finetuning on the VCG dataset. We use a larger dropout rate of 0.3 as the VCG dataset is much smaller than the entire pretraining dataset. The model converges after 30 epochs. We use a single GPU with a batch size of 64. Finetuning the model takes around 40 hours.

2 Additional Generated Examples

Table 2 and Table 3 show additional examples from our model on the VCG validation set. All the commonsense inferences are generated by the best performed model.

3 KCG Filtering Examples

Table 1 shows the average cross-entropy of our model on the generated COMET sentences. Lower cross-entropy indicates the generated inference sentences are more reasonable.

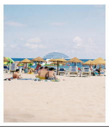
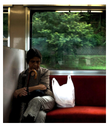
Image	Event	Task	Label	cross-entropy
	A lot of people that are at the beach	after	gets sunburned	2.755
		before	to drive to the beach	3.100
		intent	to have fun	3.398
		intent	to be safe	4.079
	Children sitting at computer stations on a long table	intent	to listen to the music	2.234
		before	to have a computer	2.847
		intent	to play with the little girl	3.710
		after	gets yelled at	4.055
	A woman is wearing a pink helmet and riding her bike through the city	intent	to get to the city	2.255
		after	gets hit by a car	2.761
		before	to buy a bike	3.052
		after	gets exercise	4.815
	A baseball player preparing to throw a pitch during a game	intent	to win the game	2.241
		after	gets hit by a ball	2.773
		before	to go to the stadium	3.222
		intent	to get a tan	4.405
	An older woman riding a train while sitting under it's window	before	to go to the train station	1.797
		intent	to get off the train	1.922
		intent	to go to the park	3.232
		after	refreshed	8.125

Table 1: Examples of commonsense descriptions generated by COMET. Examples with lower cross entropy are more reasonable. Here "Event" refers to captions in SBU and COCO dataset.

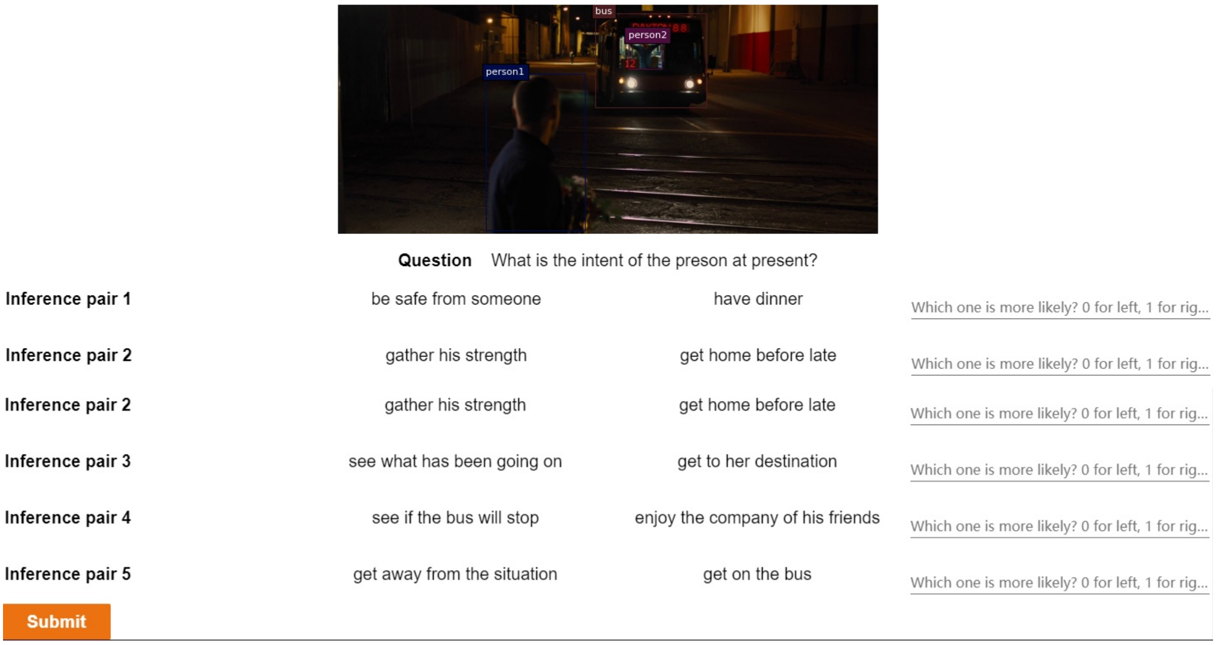
4 Additional Information on Human Evaluation

Figure 1 is the user interface of our human evaluation. We hire workers from Amazon Mechanical Turk. We reject examples with a submission time of less than 30 seconds. The median submission time is 182 seconds. We pay for each example 0.2 USD, which is around 10.4 USD per hour.

*The first three authors contribute equally to this work.

[1] https://huggingface.co/transformers/

[2] https://pytorch.org/

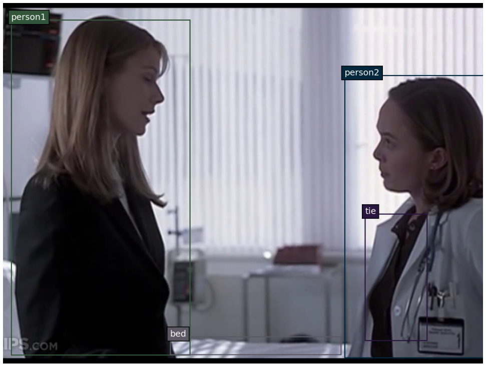


Question What is the intent of the preson at present?

Inference pair 1	be safe from someone	have dinner	Which one is more likely? 0 for left, 1 for rig...
Inference pair 2	gather his strength	get home before late	Which one is more likely? 0 for left, 1 for rig...
Inference pair 2	gather his strength	get home before late	Which one is more likely? 0 for left, 1 for rig...
Inference pair 3	see what has been going on	get to her destination	Which one is more likely? 0 for left, 1 for rig...
Inference pair 4	see if the bus will stop	enjoy the company of his friends	Which one is more likely? 0 for left, 1 for rig...
Inference pair 5	get away from the situation	get on the bus	Which one is more likely? 0 for left, 1 for rig...

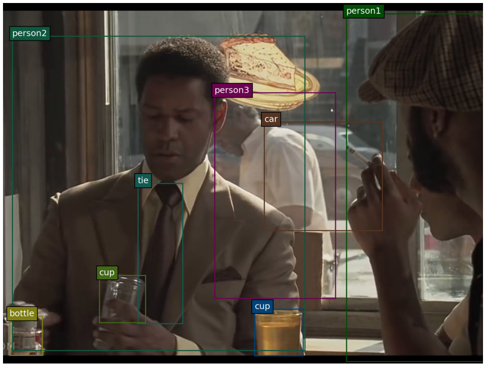
Figure 1: User interface for human evaluation.

Event: 1 is talking to 2 a doctor



Task	Ground Truth	Input	KM-BART	VCG
intent	ask 2 a question find out medical information	without event	**go home** **say goodbye to 2** **hear 2's opinion**	make herself felt better maintain the political demeanor enjoy the company of his girl friend
		with event	**hear the doctor's diagnosis** **ask the doctor some questions** **get her opinion on the procedure**	talk about her injuries heal his leg do her job as a nurse
before	feel scared for her sick relative follow 2 into an empty room see 2 go into another room	without event	**take her test results** **walk up to 2** **enter the patient's room**	decide on an outfit for the event lose a bet check his schedule to see what time it is
		with event	**meet 2 in the hospital** **walk into the room** **read a diagnosis**	call 2 into his office hear of a prescription taking visit 2 in the hospital
after	ask 2 how bad her condition is tell 2 her loved one needs help leave the hospital	without event	**leave the hospital** **walk out the door** **introduce themselves to 2**	talk about something serious with 1 greet the man walk away"
		with event	**tell 2 her symptoms** **get some medicine for 2** **ask 2 some questions**	wait patiently hug 2 listen to the response from 2

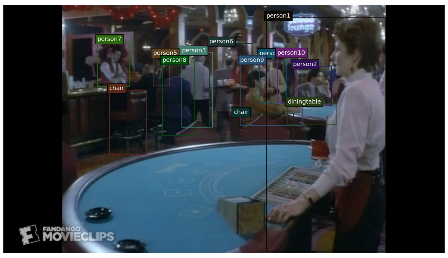
Event: 1 is sitting at the table with 2 smoking a cigarette



Task	Ground Truth	Input	KM-BART	VCG
intent	smoke a cigarette talk with 2 about something	without event	**spend quality time with 2** **stay at ease** **speak to 2**	have 1 shake hands nod in agreement do what 1 says
		with event	**have a smoke** **get to know 2 better** **get a nicotine fix**	have lunch with 2 satisfy his craving for nicotine light up
before	order food from the waiter take a drink from their water cup be seated at a table at the restaurant	without event	**order the drink** **notice 2 sitting alone at the table** **enter a restaurant**	say bye to 1 sip the drink look up from the food
		with event	**take out a cigarette** **want a light** **have a seat at the table**	have 2 meet him for dinner get a cigarette from 2 light the cigarette
after	offer to help 2 get sugar for his coffee discuss business with 2 watch 2 leave the restaurant	without event	**finish their meal** **tell 2 something important** **order lunch2**	chat while she waits for her food hug his friend watch his partner 's reaction
		with event	**finish his meal** **continue his conversation with 2** **blow out smoke**	finish smoking reminisce hand the cigarette to 2

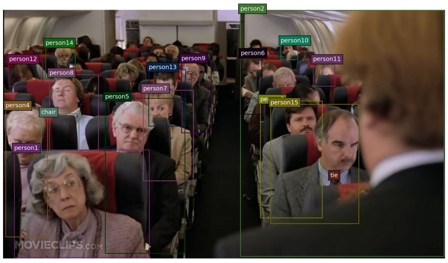
Table 2: Additional examples from the VCG validation set. Generated with nucleus sampling (top $p = 0.9$) .The bold texts are generated by KM-BART. We chose the KM-BART models which have the best performance, with or without event descriptions, respectively.

Event: 7 is a bartender serving a customer a drink



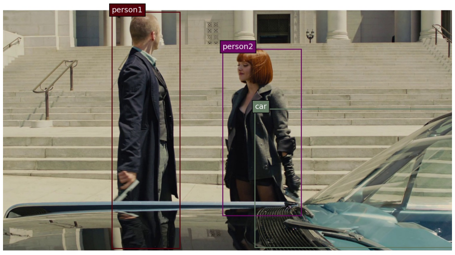
Task	Ground Truth	Input	KM-BART	VCG
intent	make the customer happy	without event	**make sure the customers were happy**	look nice for the photo
	enjoy serving others	with event	**get a good tip**	earn a good tip
before	take a customers order	without event	**get a job as a bartender**	be dressed in a suit
	walk out from behind the bar	with event	**get behind the bar**	take the customer 's money
after	bring in the drink	without event	**take the drink back to the kitchen**	walk away from the table
	ask the customer for payment	with event	**ask the customer if they want another drink**	take money from the customer

Event: 2 stand in the front of the plane and faces the passengers



Task	Ground Truth	Input	KM-BART	VCG
intent	make an announcement	without event	**ask 1 a question**	see what was happening
	tell the passengers about emergency exits	with event	**give the passengers instructions**	make sure everyone had a ticket
before	wait for the passengers to all take their seats	without event	**board the plane**	walk into the room
	walk to the front of the cabin	with event	**walk up to the front of the plane**	get on the plane
after	demonstrate how the exits work	without event	**ask 1 to sit down**	walk away from the table
	ask the passengers if they have questions	with event	**give a speech**	give the passengers a tour

Event: 1 holds the gun to his side looking up at the entrance to the building



Task	Ground Truth	Input	KM-BART	VCG
intent	scan the area for a hostile presence	without event	**get in the car**	get to the car
	be armed for a confrontation exits	with event	**be ready to shoot**	make sure no one got hurt
before	draw his weapon	without event	**walk up to 2**	walk up to the car
	drive to building to do crime	with event	**pull out his gun**	get out of the car
after	search for the person he wants to shoot	without event	**walk away from 2**	walk away
	enter building with gun	with event	**walk up to the building**	shoot at the entrance

Table 3: Additional examples from the VCG validation set. Generated with greedy search. The bold text are generated by KM-BART. We chose the KM-BART models which have the best performance, with or without event descriptions, respectively.