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

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

Abstract

We present **K**nowledge Enhanced **M**ultimodal **BART** (KM-BART), which is a Transformer-based sequence-to-sequence model capable of reasoning about commonsense knowledge from multimodal inputs of images and texts. We extend the popular BART architecture (Lewis et al., 2020) to a multi-modal model. We design a new pretraining task to improve the model performance on Visual Commonsense Generation task. Our pretraining task improves the Visual Commonsense Generation performance by leveraging knowledge from a large language model pretrained on an external knowledge graph. To the best of our knowledge, we are the first to propose a dedicated task for improving model performance on Visual Commonsense Generation. Experimental results show that by pretraining, our model reaches state-of-the-art performance on the Visual Commonsense Generation task (Park et al., 2020).

1 Introduction

"A picture is worth a thousand words" is a common adage. Indeed, a lot of human knowledge is conveyed by a combination of language and images. Scientific papers or patents usually feature both text and figures, and so do countless other data sources, e.g. math quizzes for high school students, or news reports in magazines. The combination of language and visuals is commonplace, so it is only natural to ask to what extent this combination may help machines to understand meaning.

Recently, Visual-Language (VL) tasks, including Visual Question Answering (VQA) (Antol et al., 2015), Visual Commonsense Reasoning (VCR) (Zellers et al., 2019), etc, have drawn increasing attention from the community. These VL tasks require the pretrained models to simultaneously process multimodal inputs for jointly comprehending visual and textual information. Inspired by successful pretrained language models like BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019), numerous multimodal image-text pretraining and representation learning models (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2020; Yu et al., 2020) have been proposed. These multimodal pretrained models use BERT as backbone and are denoising autoencoders, which are trained to predict the alignment of image-text segments and to predict the semantics of masked words and image regions.

To further bridge the gap between visual and textual clues in multimodal data, a model should not only comprehend cross-modal representations, but also acquire generation capabilities to complete generation tasks, for example, Image Captioning (You et al., 2016). However, applying directly a model pretrained on VL understanding tasks to generation tasks is infeasible, as these models are merely BERT-based encoders, and are thus not suitable for generation tasks. To ease this problem, researchers propose various multimodal models capable of generation tasks (Zhou et al., 2020; Li et al., 2020). These models achieve state-of-the-art performance in various downstream multimodal generation tasks, including Image Captioning (You et al., 2016), Visual Question Answering (Antol et al., 2015), etc.

Despite success in downstream multimodal generation tasks, previous models are mainly pretrained on general visual and language understanding tasks such as masked language modeling and masked region modeling. These tasks enable the models to build an alignment between visual and language features, but at the same time are inadequate to enhance the model ability in conducting complex multimodal commonsense reasoning, which requires the model to understand the underlying relations between objects. While traditionally researchers have been focusing on commonsense

*The first three authors contribute equally to this work.

reasoning on natural language (Rajani et al., 2019; Trinh and Le, 2018), recent works have paid attention to commonsense reasoning on both visual and language inputs. Zellers et al. (2019) propose the task of Visual Commonsense Reasoning (VCR), in which the model is supposed to answer multiple-choice questions about commonsense. Although VCR has pushed the limit of multimodal understanding by leveraging commonsense reasoning, the task only test the model ability on choosing instead of generating the right commonsense.

A newly introduced dataset, Visual Commonsense Generation (VCG) (Park et al., 2020), solves the aforementioned problem by requiring the model to generate commonsense texts such as cause, intention, and effect based on visual and textual inputs (see Table 5 for examples). In this work, we propose to tackle the task of Visual Commonsense Generation by leveraging our **K**nowledge Enhanced **M**ultimodal **BART** (Lewis et al., 2020), which we call **KM-BART**[1]. KM-BART is a Transformer-based model consisting of an encoder and a decoder and is pretrained on carefully designed tasks for Visual Commonsense Generation. Figure 1 presents our model architecture.

Our contributions in this work are three-folded:

1. We extend the BART model to process multimodal data of image and texts, and enable multimodal reasoning by introducing task-relevant tokens.

2. To enhance the ability on Visual Commonsense Generation, we implicitly incorporate textual commonsense knowledge in the multimodal BART model by designing a novel pretraining task. This pretraining task leverages an external language model which is designed only for textual commonsense knowledge generation.

3. We compare our pretraining task with standard pretraining tasks such as masked language modeling (MLM) and masked region modeling (MRM). Ablation studies show that our knowledge-enhanced pretraining task, despite only with textual commonsense knowledge, is more useful than standard cross-modal pretraining tasks.

[1]Our code will be available after publication.

2 Related Work

This section is organized as follows: Section 2.1 reviews previous work on Vision-Language models. We then review the literature for Commonsense Knowledge in Section 2.2.

2.1 Vision-Language Models

Early work on Vision-Language models has been largely focused on pure understanding tasks. Tan and Bansal (2019) propose LXMERT, which encodes visual feature and textual feature with a convolutional neural network and a multi-layer Transformer encoder, respectively. A cross-modality Transformer-based encoder is then used for fusing information from both visual and textual inputs. The model is pretrained on image-text matching. Lu et al. (2019) learns multimodal features by using a single unified encoder for visual and textual features. These models, although improving model performance on understanding tasks such as image text matching, are not capable of multimodal generation tasks including image captioning and visual question answering (Antol et al., 2015). Zhou et al. (2020) extends vision-language models by using a Transformer-based network as both an encoder and a decoder, making the model capable of generating texts based on visual and textual inputs. While Li et al. (2020) propose OSCAR, which improve the generation ability by introducing object tags as an additional clue during pretraining.

Previous work on multimodal pretraining for Visual-Language tasks has been largely focused on general-purpose pretraining tasks, such as masked language modeling and masked region modeling (Tan and Bansal, 2019; Lu et al., 2019; Zhou et al., 2020). These pretraining tasks focus on building semantic alignments between visual and textual features, while fall short of enabling the model for reasoning complex commonsense.

2.2 Commonsense Knowledge

Commonsense knowledge refers to the basic level of practical knowledge and reasoning about everyday situations and events commonly shared among most people (Sap et al., 2020). For example, one is supposed to know that "water is for drink" and "sunshine makes people warm". Simple as it looks, enabling artificial intelligence to conduct commonsense reasoning has been a difficult task for learning-based models (Gunning, 2018). To overcome this problem, researchers have re-

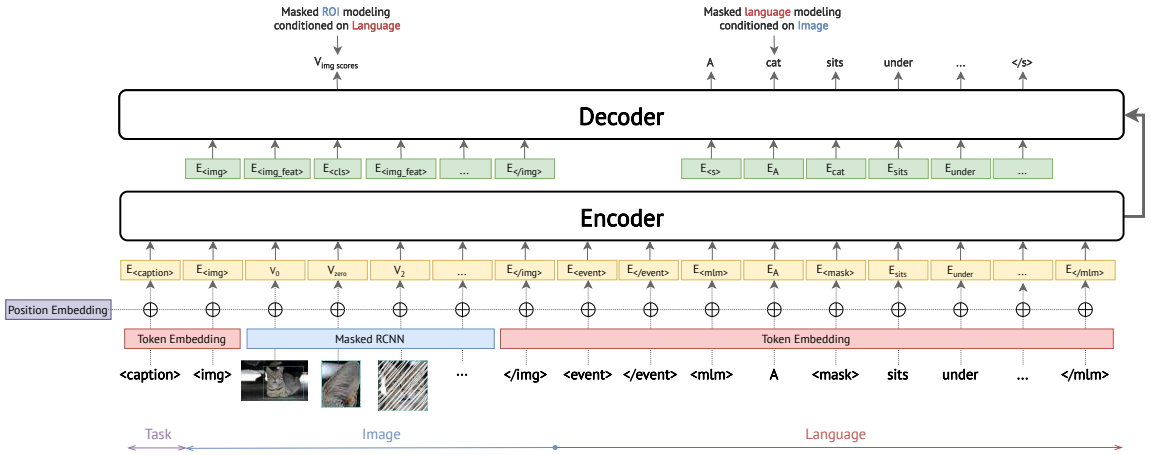

Figure 1: Model architecture. Our model is based on BART. Conditioned on prompts indicating the task type, our model is capable of generating texts based on visual and textual inputs from the Encoder. Our model uses different special tokens to indicate task types and inform the model of different modalities of input.

sorted to knowledge graphs due to their clear graph-structured representation of knowledge. For example, ConceptNet Speer et al. (2017) is a knowledge graph with nodes representing general concepts and edges indicating relational knowledge between concepts. Another commonsense knowledge graph, ATOMIC (Sap et al., 2019), extends nodes to natural language phrases, whose edges encode relations such as *intent*, *attribution*, *effect*, etc.

Despite improvements on modeling commonsense, knowledge graph-based methods suffer from the problem of heavy human engineering, resulting in difficulties in scaling. Furthermore, model performance deteriorates significantly when a node matching failure occurs (Lin et al., 2019). Recently, Bosselut et al. (2019) propose COMET, which is a Transformer-based, generative model pretrained on external commonsense knowledge graphs. Given a natural language phrase and a relation type, COMET generates natural language commonsense descriptions.

In summary, previous work for Vision-Language models does not pay enough attention to commonsense generation as the pretraining tasks are mostly general-purpose and are not particularly suitable for downstream commonsense generation task. For commonsense modeling, although researchers have proposed neural network-based generative models, such as COMET (Bosselut et al., 2019). The model is not capable of dealing with cross-modality inputs of images and texts.

In this paper, we design a pretraining task dedicated for improving model performance on commonsense reasoning tasks. We use commonsense

	#images	#sentences
Conceptual Captions (Sharma et al., 2018)	2,683,686	2,683,686
SBU (Ordonez et al., 2011)	780,750	780,750
COCO (Lin et al., 2014)	82,783	414,113
Visual Genome (Krishna et al., 2017)	86,461	4,322,358
Total	3,633,680	8,200,907

Table 1: Statistics of pretraining datasets.

knowledge induced from COMET to pretrain our proposed model. To the best of our knowledge, we are the first to apply a dedicated pretraining task for commonsense generation on cross modality visual and textual inputs. Our model is based on BART Lewis et al. (2020), which is a Transformer-based autoencoder. BART generalizes BERT as a bidirectional encoder, and GPT-2 as a left-to-right decoder. BART outperforms BERT in various downstream NLP applications and at the same time is capable of conducting generation tasks, for which BERT is not suitable. Section 3.1 delineates our model architecture and pretraining tasks.

3 Methodology

We give the details of our model architecture and pretraining tasks in this section.

3.1 Model Architecture

Figure 1 illustrates the architecture of our KM-BART. The backbone of our model is BART (Lewis et al., 2020), which is a Transformer-based sequence-to-sequence autoencoder. We modify the original BART to adapt the model to cross-modality inputs of images and texts. To adapt the model to different pretraining/evaluation tasks, we

design different special tokens. In the following subsections, we give the details of our visual feature extractor, the encoder and the decoder.

3.1.1 Visual Feature Extractor

Following previous work on Vision-Language models (Tan and Bansal, 2019; Lu et al., 2019), we use a convolution neural network pretrained on COCO dataset to extract visual embeddings, which are subsequently fed to the Transformer-based cross modality encoder of KM-BART. Specifically, we use the pretrained Masked R-CNN (He et al., 2017) from detectron2[2]. For each image, the pretrained Masked R-CNN proposes the bounding boxes for detected objects. The area within a bounding box is a **R**egion **o**f **I**nterest (RoI). We leverage the intermediate representations of the RoIs in the Masked R-CNN to obtain fixed-size visual embeddings $\{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{N-1}\} \in \mathbb{R}^d$, where N is the number of RoIs for an image and d is the dimension of the visual embeddings. For each of the RoIs, the Masked R-CNN also outputs the class distribution $p(\boldsymbol{v}_i)$, which is later used for Masked Region Modeling.

3.1.2 Encoder

Following Lewis et al. (2020), the encoder of our model is based on a multi-layer bidirectional Transformer. We add special tokens to adapt it to our pretraining and downstream evaluations tasks. Specifically, each example starts with a special token indicating the task type of the current example.

For our pretraining task of Knowledge-Based Commonsense Generation (see Section 3.2.1), we use `<before>`, `<after>`, or `<intent>` as the starting special token. For Attribution Prediction and Relation Prediction, we use `<region_caption>` (Section 3.2.2). Finally, for Masked Language Modeling and Masked Region Modeling, we use `<caption>`.

Furthermore, to inform the model of different modalities of input, we add three sets of different special tokens: For images, we use `<img>` and `</img>` to indicate the start and the end of visual embeddings, respectively.

We design different special tokens to distinguish between two sets of textual inputs: *events* and *captions*. Events are image descriptions which the model uses for reasoning about future/past events or intents of characters in the commonsense generation tasks, while captions are for Masked Language Modeling, where linguistic information plays a more important role. Hence, to inform the model of these two types of textual inputs, we use `<event>` and `</event>` for events, and `<mlm>` and `</mlm>` for captions. In the following sections, we denote embeddings for words and specical tokens by by $\mathbf{w} = \{\mathbf{E_0}, .., \mathbf{E}_{T-1}\} \in \mathbb{R}^d$, where T is the length of textual inputs.

3.1.3 Decoder

The decoder of our model is also a multi-layer Transformer. Different from the encoder, which is bidirectional, the decoder is unidirectional as it is supposed to be autoregressive when generating texts. The decoder does not take as inputs the visual embeddings. Instead, we use embeddings of the special token `<img_feat>` to replace the actual visual embeddings. For Masked Region Modeling and Masked Language Modeling, we use `<cls>` to replace the masked regions or words (see Figure 1). The model predicts the masked words and the class distribution of the masked regions during pretraining.

3.2 Pretraining Tasks

To pretrain our model, we use three image-text datasets: Conceptual Captions Dataset (Sharma et al., 2018), SBU Dataset (Ordonez et al., 2011) and Microsoft COCO Dataset (Lin et al., 2014). Statistics of the datasets are given in Table 1. The above datasets consist of examples of parallel images and texts and are widely used in previous work (Tan and Bansal, 2019; Lu et al., 2019; Zhou et al., 2020; Yu et al., 2020).

3.2.1 Knowledge-Based Commonsense Generation

The knowledge-based commonsense generation (KCG) task aims to improve KM-BART's performance in the Visual Commonsense Generation task. We leverage knowledge from COMET (Bosselut et al., 2019), which is a large language model pretrained on external commonsense knowledge graphs. Given a natural language phrase and a relation as input, COMET generates natural language phrases as commonsense descriptions. Relations of COMET include `xIntent`, `xWant`, `xNeed`, `xReact` and `xEffect`.

Due to computational limits, we only use COMET to generate new commonsense descriptions from SBU and COCO dataset. For each

[2] https://github.com/facebookresearch/detectron2

image-text pair, we use COMET to generate commonsense descriptions from the text using all five relations mentioned above. To adapt COMET generated commonsense knowledge to Visual Commonsense Generation, we consider `xIntent` and `xWant` as `intent`, `xNeed` as `before`, `xReact` and `xEffect` as `after`. In this way, we generate commonsense knowledge for SBU and COCO dataset. The newly generated dataset has more than 3.6 million examples (Table 2). However, the generated commonsense knowledge does not always make sense as only textual information is used while the image of each example is ignored. To ease this problem, we further clean the dataset by employing a self-training based data cleaning strategy.

Self-Training Based Data Cleaning The goal of our data cleaning strategy is to filter the generated commonsense knowledge dataset so that the examples in the filtered dataset closely resemble the examples in the Visual Commonsense Generation (VCG) dataset. To achieve this goal, we first initialize our KM-BART with BART parameters, and finetune KM-BART on VCG dataset for 30 epochs. The finetuned KM-BART already has a good performance on the VCG dataset, reaching a CIDEr score of 39.13, which is more than doubled compared to Park et al. (2020).

We then leverage this finetuned model to evaluate the quality of commonsense descriptions generated by COMET. We feed the corresponding image, text and relation as inputs to the finetuned KM-BART, and then compute the cross entropy loss of COMET generated commonsense descriptions. From Table 2, we can see that commonsense descriptions with a lower cross-entropy loss make more sense. Notice that when computing the cross-entropy loss of the COMET generated commonsense descriptions, our KM-BART leverages not only the textual inputs but also the visual inputs.

We compute the cross-entropy loss for all the examples in the VCG dataset as well as in the new dataset generated by COMET. Figure 2 shows the distributions of cross-entropy loss for the two datasets. We observe that commonsense descriptions generated by COMET results in higher cross-entropy losses. This is expected as images are completely ignored when using COMET to generate natural language commonsense descriptions. We only keep the examples of which cross-entropy loss is below 3.5. Table 2 shows the statistics of

	#Original	#Cleaned
SBU (Ordonez et al., 2011)	2,032,385	808,425
COCO (Lin et al., 2014)	1,653,075	660,020
Total	3,685,460	1,468,445

Table 2: Statistics of datasets before and after filtering.

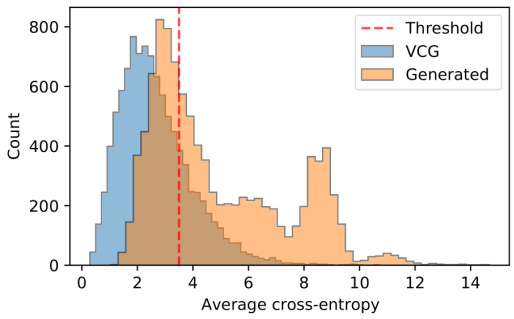

Figure 2: The distribution of the average cross-entropy on 10000 samples in the VCG dataset and our enhanced dataset. For the generated dataset, we can keep the examples of which cross entropy loss is below 3.5.

generated datasets before and after data cleaning. By cleaning, we keep only 1.46 million examples, which roughly account for 40% of the original examples.

Finally, we leverage the newly generated commonsense knowledge dataset by pretraining KM-BART on it. We expect by pretraining, the model reaches higher performance on VCG dataset. Let $\mathbf{s}^t = \{\mathbf{E}_0^t, ..., \mathbf{E}_{L-1}^t\}$ be the groundtruth commonsense of the newly generated dataset, the loss function is:

$$\mathcal{L}_{KCG}(\theta) = \\ -\mathbb{E}_{(\mathbf{E},\mathbf{v})\sim D} \sum_{i=1}^{L} \log(P_\theta(\mathbf{s}_i^t | \mathbf{s}_{<i}^t, \mathbf{w}, \mathbf{v}, t)) \quad (1)$$

where L is the length of the generated sequence, and $\boldsymbol{w}$ and $\boldsymbol{v}$ are textual inputs and visual inputs, respectively.

3.2.2 Attribute Prediction and Relation Prediction

The Visual Genome dataset consists of 2.3 million relationships and 2.8 millions attributes. To utilize these data, we propose the attribute prediction (AP) and the relation prediction (RP) pretraining tasks. The AP and RP pretraining task enable the model to learn some intrinsic properties of the scene and encode such information into the output feature vectors.

In the AP task, we feed the output vectors of the decoder for each image feature into an MLP

Event	Task type	Label	Average cross-entropy
A man with a red helmet on a small moped on a dirt road	before	to get on the bike	1.96
Children sitting at computer stations on a long table	intent	to get a snack	3.95
A girl standing with a cell phone in her hands	after	gets called a liar	5.00

Table 3: Examples of commonsense descriptions generated by COMET. We can see that examples with a lower cross entropy loss are more reasonable. Here "Event" refers to texts in SBU and COCO dataset. The corresponding images are omitted for simplicity.

classifier. In the RP task, we concatenate two output vectors of the decoder for each pair of image features and feed it into another MLP classifier. We use cross-entropy loss for both tasks.

We denote the AP indices by $0 \leq \mathbf{a} \leq A-1$, the RP indices by $0 \leq \mathbf{r} \leq R \times R$ where A is the the number of AP labels, and R is the number of RP labels. We denote the label for the AP task by $L_{\mathbf{a}}(\mathbf{v}_{\mathbf{a}}^{(i)})$, and the label for the RP task by $L_{\mathbf{r}}(\mathbf{v}_{\mathbf{r}_1}^{(i)}, \mathbf{v}_{\mathbf{r}_2}^{(i)})$. The loss function for the AP task is:

$$\mathcal{L}_{AP}(\theta) = - \mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \sum_{i=0}^{A-1} \log(P_\theta(L_{\mathbf{a}}(\mathbf{v}_{\mathbf{a}}^{(i)}) \mid \mathbf{w}, \mathbf{v})) \quad (2)$$

The loss function for the RP task is:

$$\mathcal{L}_{RP}(\theta) = - \mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \sum_{i=0}^{R\times R-1} \log(P_\theta(L_{\mathbf{r}}(\mathbf{v}_{\mathbf{r}_1}^{(i)}, \mathbf{v}_{\mathbf{r}_2}^{(i)}) \mid \mathbf{w}, \mathbf{v})) \quad (3)$$

3.2.3 Masked Language Modeling

Following Devlin et al. (2019) and Liu et al. (2019), we randomly mask the textual description with a probability of 15% in masked language modeling (MLM) task. Within this 15% of text, we use <mask> to replace the masked word with a probability of 80%, use a random word to replace with a probability of 10% and keep the masked word unchanged with a probability of 10%.

We denote the mask indices by $0 \leq \mathbf{m} \leq M-1$, where M is the number of masked words. We denote the remaining words that are not masked by $\mathbf{w}_{\backslash \mathbf{m}}$, the loss function is:

$$\mathcal{L}_{MLM}(\theta) = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \log(P_\theta(\mathbf{w}_{\mathbf{m}} | \mathbf{w}_{\backslash \mathbf{m}}, \mathbf{v})) \quad (4)$$

3.2.4 Masked Region Modeling

In masked region modeling (MRM) task, we sample image regions and mask the corresponding feature vectors with a probability of 15%. The masked vector will be replaced by a vector filled with zeros. The model needs to predict the distribution over semantic classes for the masked regions. The loss function is to minimize the KL divergence of the output distribution and the distribution predicted by the Masked R-CNN used in visual features extraction.

We denote the mask indices by $0 \leq \mathbf{m} \leq M-1$, where M is the number of masked regions. We let $p(\mathbf{v}_{\mathbf{m}}^{(i)})$ denote the class distribution detected by Masked R-CNN, $q_\theta(\mathbf{v}_{\mathbf{m}}^{(i)})$ denote the class distribution output by our model, the loss function is:

$$\mathcal{L}_{MRM}(\theta) = \mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \sum_{i=1}^{M} D_{KL}(p(\mathbf{v}_{\mathbf{m}}^{(i)}) || q_\theta(\mathbf{v}_{\mathbf{m}}^{(i)})) \quad (5)$$

3.2.5 Combining Losses

To combine all the loss we described above, we sum the loss weighted by $W_{KCG}, W_{AP}, W_{RP}, W_{MLM}, W_{MRM} \in \mathbb{R}$. The weights are chosen to roughly balance every term during the training phase. The final loss is:

$$\mathcal{L} = W_{KCG}\mathcal{L}_{KCG} + W_{AP}\mathcal{L}_{AP} + W_{RP}\mathcal{L}_{RP} + W_{MLM}\mathcal{L}_{MLM} + W_{MRM}\mathcal{L}_{MRM} \quad (6)$$

Notice that during the pretraining, KCG and MLM cannot be done at the same time. Hence, we set W_{KCG} and W_{MLM} to zero alternatingly between consecutive batches.

4 Experiments

To evaluate our model and the pretraining tasks, we designed multiple experiments.

4.1 Settings

In our experiments, following BART-base (Lewis et al., 2020), we fix the model architecture to a 6-layers encoder and a 6-layer decoder. To understand how each pretraining task helps model performance on the downstream task of Visual Commonsense Generation, We ablate on pretraining tasks.

	With Event	BLEU-2	METEOR	CIDEr
Park et al. (2020)	Y	13.50	11.55	18.27
Random init				
w/o pretraining	Y	22.28	14.55	36.49
+ KCG	Y	22.16	14.52	37.06
+ KCG (before filtering)	Y	22.24	14.43	37.08
+ AP & RP	Y	22.49	14.64	37.18
+ MLM & MRM	Y	22.44	14.70	37.44
BART init				
w/o pretraining	Y	22.86	**15.17**	39.13
+ KCG	Y	**23.47**	15.02	**39.76**
+ KCG (before filtering)	Y	22.90	14.98	39.01
+ AP & RP	Y	22.93	14.99	39.18
+ MLM & MRM	Y	23.01	14.91	38.12
Park et al. (2020)	N	10.21	10.66	11.86
Random init				
w/o pretraining	N	13.54	10.14	14.87
+ KCG	N	13.46	10.09	14.97
+ KCG (before filtering)	N	13.34	9.914	14.48
+ AP & RP	N	13.83	10.28	15.48
+ MLM & MRM	N	14.36	10.73	16.72
BART init				
w/o pretraining	N	11.68	10.09	11.33
+ KCG	N	7.599	8.692	5.787
+ KCG (before filtering)	N	13.08	9.898	13.58
+ AP & RP	N	7.861	8.800	5.995
+ MLM & MRM	N	13.78	10.55	15.03

Table 4: Results on VCG validation set.

We use the following experimental settings: (1) Without any pretraining; (2) Only with knowledge-based commonsense generation; (3) Only with attribute prediction and relation prediction; (4) Only with masked language modeling and masked region modeling.

For each of the above settings, we initialize the model from random or from BART weights, respectively. Besides, we also test the model performance when only using images as inputs. All the models are pretrained for 20 epochs on 4 Titan RTX GPUs with an effective batch size of 256. We use the Adam optimizer with a learning rate of 1e-5. We use the default setting for all other parameters of the Adam optimizer. To speed up training, we use PyTorch built-in automatic mixed precision.

4.2 Evaluation Task

We evaluate our model on the recently proposed Visual Commonsense Generation (VCG) Dataset (Park et al., 2020). Given an image and a description of the event in the image, the task aims to predict events happen before/after, and the present intents of the characters in the given image. Table 5 gives examples of the dataset.

The dataset consists of 1174K training examples and 146K validation examples. Some examples in the dataset share the same images or events, but with different inferences for events before/after or intents at present. Table 5 gives some examples of the dataset. We report our model performance on the validation set as the test set is not available yet.

We finetune our pretrained model for 30 epochs on the VCG dataset. We keep the same optimizer and batch size as in pretraining. For finetuning, a drop out rate of 0.5 is used. During generating, we use a beam size of 1.

4.3 Results

Table 4 shows that our KM-BART achieves state-of-the-art performance on the VCG dataset. KM-BART outperforms Park et al. (2020) by a large margin on all the metrics. The advantage of our Knowledge-Based Commonsense Generation pretraining task is most evident when we initialize the model from BART parameters and give event descriptions when finetuning. Under this setting, the model reaches a CIDEr score of 39.76, which is more than doubled compared to 18.27 from Park et al. (2020). In the meantime, the model also manages to improve the BLEU score by almost 10 points and the METEOR score by more than 3 points.

By conducting the ablation studies, we can conclude that all four pretraining tasks improve performance. We observe that the model performance benefits more from the pretraining tasks when event descriptions are not given. We conjecture that the reasons for this is that the pretraining tasks mostly improve model ability in understanding visual features from cross-modality inputs, while the tasks with event descriptions rely more on textual information instead of visual information.

We also observe that although filtering on the commonsense generation pretraining task reduces the dataset size by more than 60%, pretraining with KCG still outperforms pretraining with KCG (before filtering) in almost all the settings. This demonstrates that our self-training based filtering technique is helpful, as it helps the model reach even higher performance with less training data. The only exception is when we finetune without event descriptions from BART parameters. In this case, we notice that model performance with KCG is significantly lower than model performance with KCG (before filtering). We argue that this results from that models initialized with BART parameters are hard to optimize as these models pays too much attention to textual inputs and have not seen visual inputs at all before finetuning.

Furthermore, the KM-BART initialized from the BART model makes a great improvement on the

Event and image	Task type	Model type	Generated Commonsense
4 is a nun standing with her arms crossed speaking to a group of boys 	intent	without event	teach the students
		with event	give the boys some advice
		ground truth	ask the boys questions figure out who misbehaved
	before	without event	walk into the classroom
		with event	gather the boys in the classroom
		ground truth	be informed the boys had done something call the boys into the office want to reprimand the boys learn that the boys did something bad
	after	without event	teach the students
		with event	give the boys detention
		ground truth	show the boys the evidence she has scold the boys punish the boys give the boys detention
1 is quickly walking down the sidewalk with a sandwich in her hand 	intent	without event	get to her destination quickly
		with event	get to work on time
		ground truth	get to her destination quickly save time by eating lunch while walking
	before	without event	walk out of the building
		with event	pick up the sandwich
		ground truth	purchase the sandwich at a deli take the subway to her current location be late for work be eating a sandwich while walking to work
	after	without event	walk into a building
		with event	eat the sandwich
		ground truth	cross the street at a crosswalk throw away the trash from the sandwich look down at the sandwich thinking it's terrible finish eating the sandwich anyway

Table 5: Two examples from the VCG dataset. We compare the the model outputs with event as input, the model without event as input and the ground truth. In the first example, the model without event recognizes that the people on the left are students, but didn't describe the group more precisely. The model with events as input receives hints from the event, and it always refers the group as "boys". In the second example, the model without event notices that the woman is walking on the street in a hurry, but it fails to recognize the blurred sandwich.

task with events, but on the task without event, it is even worse than random initialization. This is reasonable, as the BART model is only trained on textual data and initializing from BART parameters acts as a bad starting point for the task with only visual inputs.

4.4 Case Study

In Table 5, we show two examples and compare the results of our model predictions to the ground truths, which are labeled by humans. The generated sentences from the model without event can already capture the most important information of commonsense. We also observe that adding event descriptions to the inputs helps the model generate more details instead of merely generating general descriptions when only conditioned on images.

5 Conclusion and Future Work

In this paper, we propose **K**nowledge Enhanced **M**ultimodal **BART** (**KM-BART**), which is a Transformer-based model capable of reasoning about and generating commonsense descriptions from cross modality inputs of images and texts. We propose the pretraining task of Knowledge-Based Commonsense Generation, which improves the reasoning ability of KM-BART by leveraging a large language model pretrained on external commonsense knowledge graphs. We use the self-training technique to clean the automatically generated commonsense descriptions. Experimental results on downstream Visual Commonsense Generation task show that our KM-BART reaches state-of-the-art performance, and that our proposed commonsense generation pretraining task is effective, helping the model reaching comparable performance when compared to models pretrained with other general-purpose pretraining methods.

In the future, we plan to pretrain our KM-BART on all pretraining tasks of Section 3.2. We also plan to further expand our pretraining dataset for Visual Commonsense Generation by including the Conceptual Captions Dataset (Sharma et al., 2018). To further understand the advantages and limitations of our KM-BART, we will also conduct a human evaluation on the generated commonsense descriptions.

Acknowledgement We thank Prof. Mrinmaya Sachan (`mrinmaya.sachan@inf.ethz.ch`) for insightful discussions.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *ICCV*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *ECCV*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

David Gunning. 2018. Machine common sense concept paper. *arXiv preprint arXiv:1810.07528*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *ICCV*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *ACL*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *ACL*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP*.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *AAAI*.