Appendix for K-VQG: Knowledge-aware Visual Question Generation for Common-sense Acquisition

1. Details of Target Knowledge Parser

Following K-VQG model, we used a model that consists of a UNITER-based encoder [2] and BART-based decoder [3] as our Target Knowledge Parser model. The encoder takes the visual embeddings $\boldsymbol{v}$ and the tokenized question $\boldsymbol{q}$. We used region features obtained from Faster R-CNN [1] as visual embeddings, as in our VQG model. The question is tokenized into input sequences using WordPiece tokenizer [4].

Our model is trained to minimizing the negative conditional log-likelihood loss function can be expressed through the following equation:

$$L = -\sum_{n=1}^{|\boldsymbol{k}|} \log P_\theta(\boldsymbol{k}_n \,|\, \boldsymbol{k}_{<n}, \boldsymbol{h}_t) \tag{1}$$

where $\boldsymbol{h}_t = \mathrm{Enc}(\boldsymbol{v}, \boldsymbol{q})$, and $\boldsymbol{k} = \{\boldsymbol{w}_h, w_{[\mathrm{SEP}]}, \boldsymbol{w}_r, w_{[\mathrm{SEP}]}, \boldsymbol{w}_t\}$. is a special token that indicates the separation of each part, and $\boldsymbol{w}_h$, $\boldsymbol{w}_r$, $\boldsymbol{w}_t$, $w_{[\mathrm{SEP}]}$ denote the tokens of the head, relation, tail phrases and special token, respectively.

2. Additional Examples of the K-VQG Dataset

We show additional examples of the K-VQG dataset below.

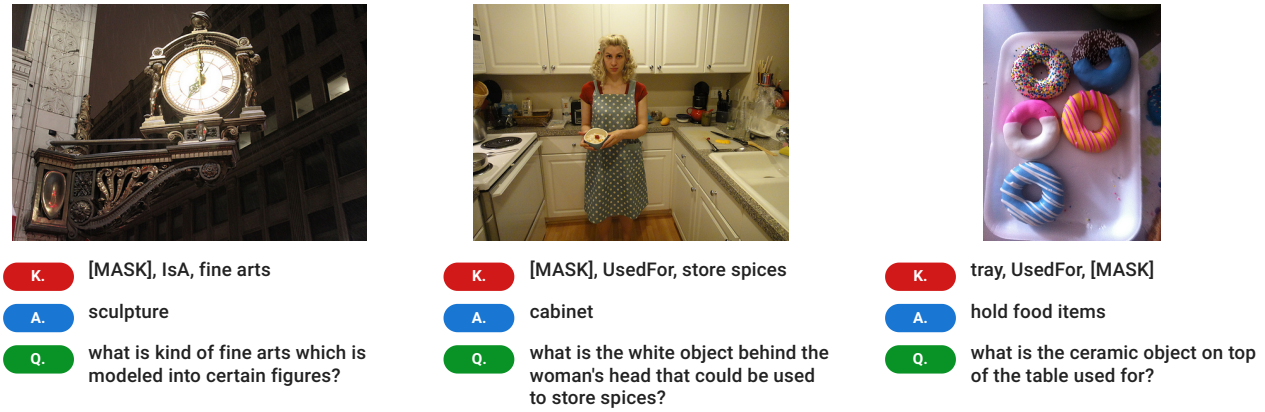


Figure 1. Additional examples of the K-VQG dataset (1)

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

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

[3] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan 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*, July 2020.

[4] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

K. [MASK], CreatedBy, seed

A. plant

Q. what is the name of the object to the right of the fruit that can be grown from seed?

K. ski, HasSubEvent, [MASK]

A. hit slopes

Q. what do you do with the footwear the man is wearing?

K. [MASK], IsA, sports shirt

A. jersey

Q. what is the sports shirt worn by the tennis player?

K. [MASK], CreatedBy, baker

A. bread

Q. what is the object called that is created by a baker and sitting on top of the bowl?

K. [MASK], DefinedAs, part of object designed to grasped by hand

A. handle

Q. what is the black shiny item that is designed to be grasped by the hand and is inside a shoe?

K. [MASK], AtLocation, shopping mall

A. bag

Q. what kind of an object is carried to the shopping mall for purchase?

K. [MASK], MadeUpOf, cheese

A. pizza

Q. what is in the tray and is made up of cheese?

K. [MASK], UsedFor, soak in

A. bathtub

Q. what object against the wall can fill with water to soak in?

K. tire, MadeUpOf, [MASK]

A. rubber

Q. what is the black outside of a tire made of?

K. [MASK], CapableOf, hunt rabbit

A. bear

Q. which animal have a capable of to hunt rabbit for their food?

K. [MASK], DefinedAs, tallest land animal

A. giraffe

Q. what is the animal standing in the grass that is defined as the tallest land animal?

K. [MASK], Desires, water and sun

A. plant

Q. what is the object that needs water and sun which is against the wooden wall?

Figure 2. Additional examples of the K-VQG dataset (2)

K. kite, AtLocation, [MASK]

A. park

Q. where do you traditionally play with the toy the kid is holding?

K. [MASK], HasProperty, yellow

A. banana

Q. what is the yellow fruit on the right called?

K. [MASK], HasA, nose

A. elephant

Q. what is the animal standing near the fence that has a long nose?

K. ski, HasPrerequisite, [MASK]

A. go to ski mountain

Q. what do you need when you go to ski mountain?

K. flag, CapableOf, [MASK]

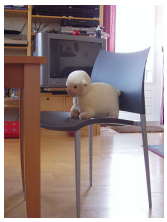
A. wave from pole

Q. what can the row of colorful objects do when hanging outside?

K. [MASK], UsedFor, sit down on

A. bench

Q. what flat wooden surface next to the table can people sit down on?



K. [MASK], IsA, device

A. television

Q. what electronic device is in the wooden entertainment center?

K. [MASK], IsA, breakfast food

A. doughnut

Q. what is the dark round object that is often ate at breakfast time?

K. [MASK], HasProperty, long neck

A. giraffe

Q. what is the large animal with the long tall neck?

K. hat, UsedFor, [MASK]

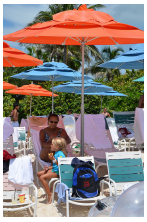
A. prottecting head

Q. what is the red thing the man is wearing used for?

K. [MASK], ReceivesAction, served in bowl

A. soup

Q. which item, served in bowl, is next to the roll?



K. [MASK], CapableOf, shade people from sun

A. umbrella

Q. what large red item on the metal pole is helping to shade people from sun?

Figure 3. Additional examples of the K-VQG dataset (3)