Uncertainty-based Visual Question Answering: Estimating Semantic Inconsistency between Image and Knowledge Base

Anonymous ACL submission

Abstract

Knowledge-based visual question answering (KVQA) task aims to answer questions that require additional external knowledge as well as an understanding of images and questions. Recent studies on KVQA inject an external knowledge in a multi-modal form, and as more knowledge is used, irrelevant information may be added and can confuse the question answering. In order to properly use the knowledge, this study proposes the following: 1) We introduce a novel semantic inconsistency measure using caption uncertainty and semantic similarity. 2) We suggest a new external knowledge assimilation method based on the semantic inconsistency measure and apply it to integrate explicit knowledge and implicit knowledge for KVQA. 3) The proposed method is evaluated on the OK-VQA dataset and achieves the state-of-the-art performance.

1 Introduction

Knowledge-based visual question answering (KVQA) task is to answer questions that require an understanding of images, questions, and additional external knowledge. The KVQA task is proposed with the aim of reaching human-level reasoning. Injecting huge knowledge based on the entities identified from images and questions in a multi-modal form is among the tasks being researched. However, as the knowledge base (KB) is often incomplete, when the context of the entities is conflicting with the KB, irrelevant information can be retrieved and confuse the question answering.

For the example in Fig 1, the question can be answered with a full understanding of the image and question. However, the predicted answer is yellow when we add the related external knowledge (i.e., the color of bananas is yellow) for KVQA. Since adding the knowledge confuses the prediction of the answer, it is necessary to adjust the amount of the external knowledge injected according to some semantic inconsistency measure. This study proposes a new approach for measuring such inconsistencies and introduces an external knowledge assimilation method. This study is summarized as follows:

Q: What color is the banana?
A: Green

Figure 1: Representative visual question answering example occurring semantic inconsistency with an external knowledge

- We introduce a new semantic inconsistency measure based on caption generation, which is an ensemble of a) uncertainty of the caption and b) similarity between the caption generated with the KB and the ground-truth caption
- We propose an external knowledge assimilation method based on the proposed semantic inconsistency measure to control the use of external knowledge in KVQA.
- We apply the proposed method for combining explicit and implicit knowledge in KVQA and achieve the state-of-the-art result when evaluated with the OK-VQA dataset.

2 Related work

2.1 KVQA approaches using pre-trained model

A lot of researches have studied image and text as a multi-modal form. By tokenizing the object in an image, a study in which alignment between an object and text has been proposed to apply a self-attention model (Lu et al., 2019; Li et al., 2019).

In addition, (Li et al., 2019) showed high performance in various downstream tasks compared with other vision-language approaches (Singh et al., 2020). Therefore, this study experiments with the approach suggested by (Li et al., 2019) for extracting the implicit knowledge. Multi-modal approaches using image features from Faster R-CNN or ResNet and question embedding of pre-trained models are also proposed (Kim et al., 2018; Ben-Younes et al., 2017). (Kim et al., 2018) generated joint representation through Bilinear Attention Map. (Ben-Younes et al., 2017) extracted image-text joint representation by using image features and question embedding and proposed a 3-way Tucker fusion method. In addition to using pre-trained models, there have also been studies trying to solve VQA tasks using additional external knowledge. (Marino et al., 2019) proposed ArticleNet using Wikipedia search API related to keywords of an image and words of the question. A method for extracting external knowledge related to the objects in an image was also introduced (Zhang et al., 2021). (Zhang et al., 2021) extracted knowledge by using the object label output from the Faster R-CNN model. This study extracts more relevant knowledge by using not only image object keywords, but also words in the question.

2.2 Graph-based KVQA approaches

Besides using pre-trained models, studies using graph-based models were proposed (Hudson and Manning, 2019; Jiang and Han, 2020; García and Nakashima, 2020; Gao et al., 2020; Ziaeefard and Lécué, 2020). (Hudson and Manning, 2019) suggested the Neural State Machine based on a probabilistic graph for reasoning on VQA. (García and Nakashima, 2020) introduced a video scene graph and caption generation method, and applied them for reasoning video question answering (video-QA) task. (Jiang and Han, 2020) studied a heterogeneous graph alignment network considering inter- and intra-modality on the video-QA. (Gao et al., 2020) proposed a method to create graphs from visual, linguistic, and numeric features and suggested an aggregator that combines the features. However, because the study considers the contents of the image, the method has a limitation on answering a question that requires additional knowledge. (Ziaeefard and Lécué, 2020) suggested graph-based VQA for capturing the interrelationship between objects and entities of external knowledge by combining concept graph and scene graph. However, the scene graph relation is limited because only locational information is considered. Therefore, in the OK-VQA dataset, the location-based scene graph extraction methods do not show significant performance.

Moreover, studies using a pre-trained model and graph-based model have been suggested (Saqur and Narasimhan, 2020; Li et al., 2020; Marino et al., 2021). (Saqur and Narasimhan, 2020) introduced multimodal graph networks for compositional generalization in VQA, but the method is evaluated on the task that requires object detection or recognition. (Li et al., 2020) proposed a Knowledge Graph Augmented model using a pre-trained model and graph-based method. However, the knowledge subgraph is generated by using the image object labels and the words of the question without the image-question context. (Marino et al., 2021) proposed to integrate image-text representation from the BERT-based model and graph information based on the concept of image objects and questions. However, when there are conflicts between graph and pre-trained model representation, there are limitations in using knowledge for question answering.

This study proposes a new approach that measures semantic inconsistencies between KB and the given problem, and moderate the use of knowledge based on the measurement.

3 Approach

This section introduces a semantic inconsistency measure that makes use of uncertainty and semantic similarity modeling.

3.1 Semantic inconsistency between an image and an external KB

In this study, we utilize caption generation to measure semantic inconsistency between an image and external KB. Inspired by (Xiao and Wang, 2021), we adopt uncertainty model of caption generation and introduce a novel measure for estimating semantic inconsistency between the KB and the VQA context.

3.1.1 Ensemble-based uncertainty estimation for KVQA

In the existing image captioning, to generate a sentence y when an input x is given, the conditional distribution $p(y|x)$ is learned and continuously predicts tokens through an autoregressive distribution.

$$p(y|x) = p(y_1|x)\prod_{i=2}^{k} p(y_i|x, y_1, \cdots, y_{i-1}) \quad (1)$$

In Eq. (1), y_i denotes the token corresponding to the index i in sentence y, and the given set $x, y_1, \cdots, y_{i-1}$ denotes context c_i for predicting the token corresponding to i. The number of tokens that can be predicted is limited based on the given context. For example, the word "beach" cannot be generated when an image of a cat on a desk is given. When a set of words irrelevant to the context is denoted hallucinated word $V_h^{(c_i)}$, the following equation can be written

$$p(y_i \in V_h^{(c_i)}) = \sum_{v \in V_h^{(c_i)}} p(y_i = v|c_i) \quad (2)$$

In image captioning, token prediction in a given context is calculated with the following cross-entropy equation. The equation can be divided into two based on an entropy of the set of words relevant to the context and that of the set of words irrelevant to the context as

$$\begin{aligned} H(y_i|c_i) &= -\sum_{v \in V} p(y_i = v|c_i) log p(y_i = v|c_i) \\ &= -\sum_{v \in V \setminus V_h^{(c_i)}} p(y_i = v|c_i) log p(y_i = v|c_i) \\ &\quad -\sum_{v \in V_h^{(c_i)}} p(y_i = v|c_i) log p(y_i = v|c_i) \end{aligned} \quad (3)$$

The uncertainty that can be predicted by the Eq. (3) can be divided into two. 1) uncertainty that appears in selection of a token that describes the context 2) uncertainty that appears due to the interference of words irrelevant to the context or an insufficient training system. The latter is directly related to calculating hallucinated words that are probability irrelevant to the given context. We make use of the latter in measuring uncertainty in KVQA, as described below. The latter can be decomposed into two: aleatoric uncertainty and epistemic uncertainty (Kiureghian and Ditlevsen, 2009; Depeweg et al., 2018; Kendall and Gal, 2017). The uncertainties can be measured by an ensemble-based model (Lakshminarayanan et al., 2017). In Eq. (4), w denotes the model weights and $q(w)$ denotes the posterior distribution of weights in the training data. If the weights are fixed, $H(y_i|c_i, w)$ represents the uncertainty related to the data. Aleatoric uncertainty can be written as $\mathbb{E}_{q(w)}[H(y_i|c_i, w)]$ and calculated by the mean of $H_m(y_i|c_i)$. Epistemic uncertainty can also be written by the difference between the entropy $H(y_i|c_i)$ of $p(y_i|c_i)$ and aleatoric uncertainty in Eq. (5). The proposed method is illustrated in Fig 2.

$$\begin{aligned} u_{al}(y_i|c_i) &= \mathbb{E}_{q(w)}[H(y_i|c_i, w)] \\ &= \frac{1}{M}\sum_{m=1}^{M} H_m(y_i|c_i) \end{aligned} \quad (4)$$

$$\begin{aligned} u_{ep}(y_i|c_i) &= H(y_i|c_i) - \mathbb{E}_{q(w)}[H(y_i|c_i, w)] \\ &= H(y_i|c_i) - u_{al}(y_i|c_i) \end{aligned} \quad (5)$$

A recent study shows that the model pre-trained with a large amount of image captioning data incorporates commonsense knowledge that is implicit in the data (Su et al., 2020). We use such a pre-trained model (with implicit commonsense knowledge) to generate image captions from the KVQA image data, and predict the uncertainty of the knowledge for the given VQA using the above ensemble model.

3.1.2 Measuring similarity between caption sentences

In addition to the above uncertainty model, this study additionally proposes a novel measure that predicts the uncertainty of the implicit knowledge based on the similarity between the generated and the ground-truth caption. That is, if the generated caption with the pre-trained model is much different from the ground-truth caption, the implicit commonsense knowledge may be not much of use for the given problem. The S-BERT sentence embedding method (Reimers and Gurevych, 2019) is used to calculate the caption similarity. The similarity between the caption embeddings is calculated as follows

$$Cap\_Sim(S_g, S_t) = \frac{f(S_g) \cdot f(S_t)}{\|f(S_g)\| \cdot \|f(S_t)\|}, f : encoder \quad (6)$$

3.2 Knowledge-based visual question answering

Based on the above uncertainty measures, we present a new approach that integrates implicit knowledge and explicit knowledge (external KB) into KVQA.

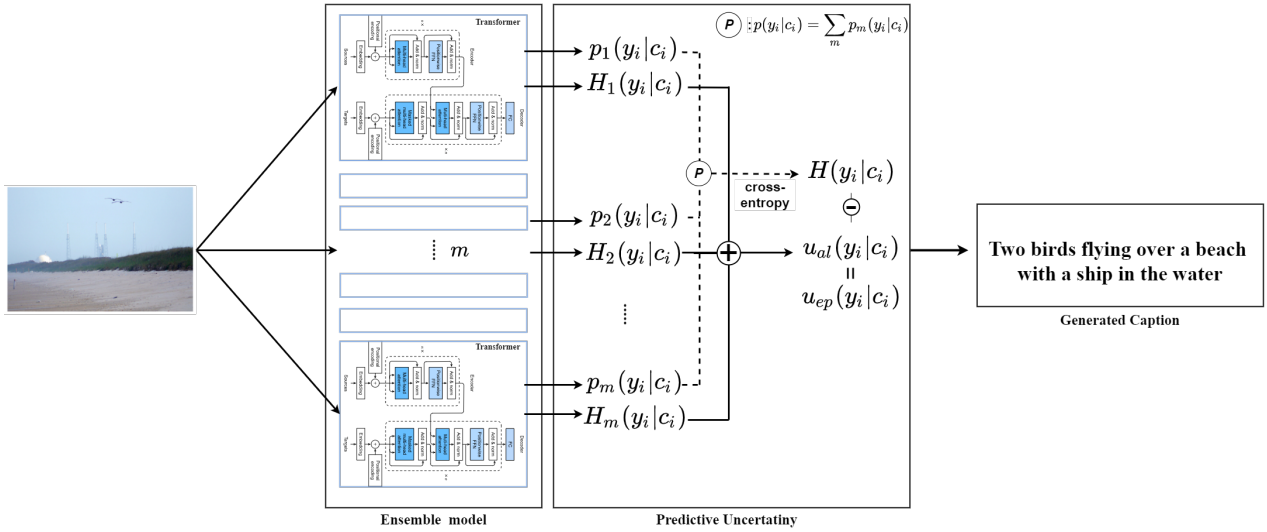

Figure 2: Ensemble-based uncertainty estimation based on caption generation

3.2.1 Use of knowledge based on semantic consistency

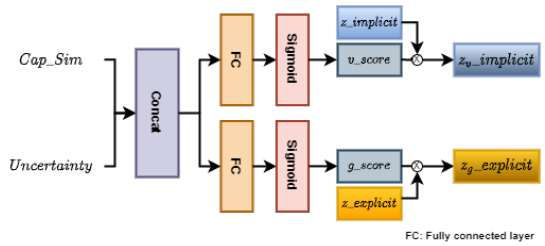

Figure 3: Use of knowledge adjusted based on uncertainty measures

As shown in Fig 3, we adjust the use of the given KB based on the above mentioned uncertainty measures. In KVQA, when the uncertainty is high and the similarity is low, the system tends to attend to the content of image-question information. Otherwise, the external knowledge is more attended.

$$
\begin{aligned}
v\_score &= \sigma(W_v * [cap\_sim, uncertainty]) \\
g\_score &= \sigma(W_g * [cap\_sim, uncertainty]) \\
z_v\_implicit &= v\_score * z\_implicit \\
z_g\_explicit &= g\_score * z\_explicit
\end{aligned} \tag{7}
$$

As shown in Fig 3 and Eq. (7), v_score and g_score are calculated through fully connected layer and sigmoid function after concatenating the similarity and the uncertainty. Each of the score values is finally represented by $z_v\_implicit$ and $z_g\_explicit$ through dot product between the pre-calculated $z\_implicit$ extracted from a vision-language model and $z\_explicit$ knowledge representation extracted from a knowledge graph. The use of image-question information and KB are adjusted based on the score.

3.2.2 KVQA with semantic consistency model

For KVQA, this study proposes a semantic consistency model that relies on the uncertainty measures described above. The model relies on two types of knowledge sources inspired by (Marino et al., 2021): 1) explicit knowledge and 2) implicit knowledge. The former is the knowledge extracted from Relational Graph Convolution Networks (RGCN) that has the external KB as input (Schlichtkrull et al., 2018). The latter is a vision-language embedding extracted from VisualBERT trained with a large-scale data. Furthermore, the use of the explicit and implicit knowledge is adjusted based on the uncertainty estimation, as described in section 3.1.

Explicit knowledge extraction: Explicit knowledge is created by extracting relevant knowledge from the external KB, using the objects recognized in the image. In this study, about 4000 image keywords including objects, places, and attributes of objects are extracted with the following models: 1) ResNet-152 (ImageNet (Deng et al., 2009)); 2) ResNet-18 (Place365 (López-Cifuentes et al., 2020)); 3) Faster R-CNN (VisualGenome (Krishna et al., 2017a)); 4) Mask-RCNN (LVIS (Gupta et al., 2019)). External KB used are as follows: 1) DBPedia (categorical information) (Auer et al., 2007);

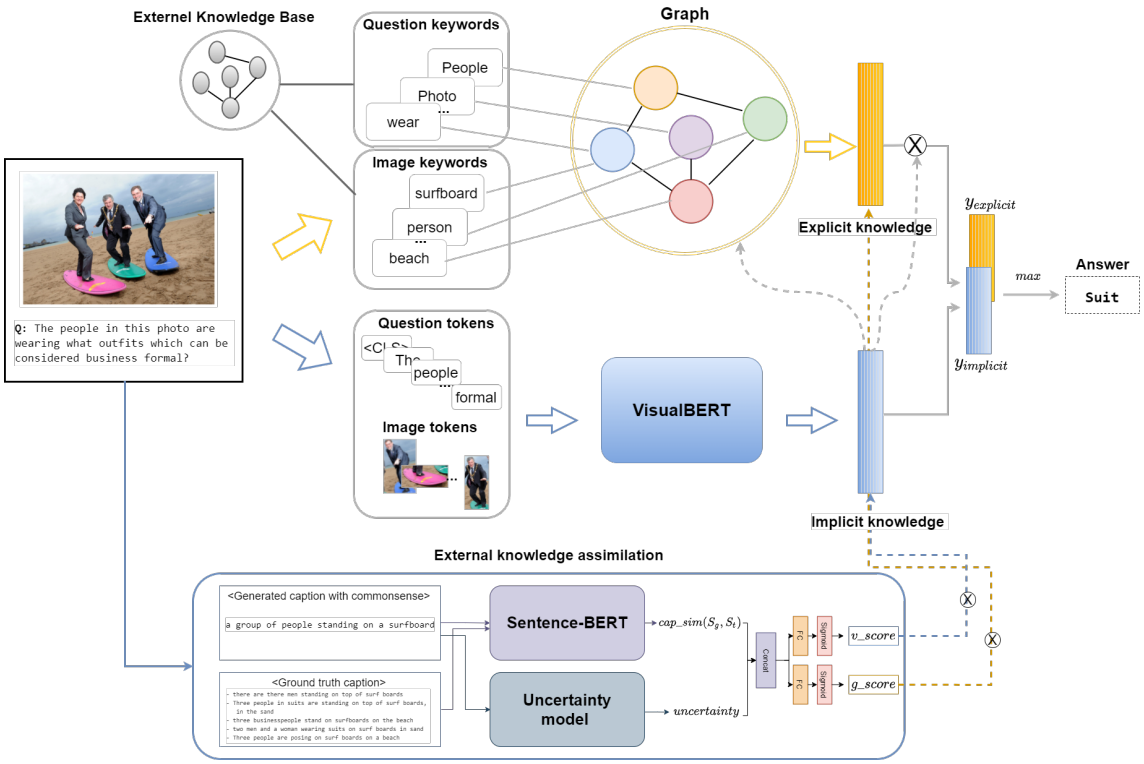

Figure 4: Overall model architecture

2) ConceptNet (commonsense knowledge)(Liu and Singh, 2004); 3) VisualGenome (spatial relationship) (Krishna et al., 2017b); 4) hasPartKB (part relationship) (Bhakthavatsalam et al., 2020). The relevant knowledge is retrieved with image keywords and question words. As a result, a total of 36,000 edges and 8,000 nodes are extracted. For integrating knowledge graphs, we use RGCN that distinguishes types and directions of edges in this study. The followings are used as RGCN inputs: 1) keyword presence that indicates words in the question with filtered words (with one-hot matrix); 2) an image keyword probability extracted from a pre-trained model; 3) Word2vec representation of each keyword or average Word2vec representation of multiple words (Mikolov et al., 2013); 4) Implicit knowledge representation $z\_implicit$ extracted from VisualBERT. The extracted explicit and implicit knowledge are integrated into KVQA as described above.

Implicit knowledge extraction: Transformer-based language models trained with a large-scale corpus are known to learn commonsense. Therefore, we use VisualBERT model to make use of the implicit knowledge generated from the image and the question (Li et al., 2019), as shown in Fig 4. Although there are various studies that align images and sentences together, we apply the appropriate model to our task by (Singh et al., 2020) experiments. The question representations are extracted by the pre-trained BERT model with BookCorpus dataset and English Wikipedia, and we use the representations as the input to the VisualBERT model. Furthermore, the visual representations are extracted from the Faster R-CNN model pre-trained with VisualGenome/COCO dataset and the result becomes the VisualBERT's input. To produce $z\_implicit$ representation, we use mean-pooling with outputs extracted from the VisualBERT model.

The final implicit and explicit knowledge representations are calculated to predict the answer from the set of answer vocabulary $V \in \mathbb{R}^v$ as follows

$$y^{implicit} = \sigma(W_v * z_v\_implicit + b) \quad (8)$$

$$y_i^{explicit} = \sigma((W_{ge} * z_{g}^{i}\_explicit + b_{ge})^T (W_{vi} * z_v\_implicit + b_{vi})) \quad (9)$$

In this study, the answer is predicted through the hidden state of word i corresponding to $V \in \mathbb{R}^v$ from the extracted explicit knowledge in Eq. (9). The final answer is selected by choosing the highest value from both $y^{implicit}$ and $y^{explicit}$. The model is trained with binary cross-entropy.

4 Experiments and results

4.1 Dataset and baseline

We use OK-VQA dataset (Marino et al., 2019) which is a popular KVQA benchmark dataset. The dataset consists of a total of 14,031 images and 14,055 questions.

Dataset	# of images	# of captions
Train	82,783	413,915
Validation	40,504	202,520
Test	40,775	379,249

Table 1: Table of MSCOCO dataset

MSCOCO dataset (Chen et al., 2015) is used to pre-train baseline models that generate captions. The dataset size is shown in Table 1. In addition, Att2in (Rennie et al., 2017), BuDn (Anderson et al., 2018), and Transformer (Vaswani et al., 2017) are selected as the baseline models for caption generation, which are the representative image captioning models, and used to generate captions of the OK-VQA dataset.

4.2 Metrics

In this study, a standard evaluation metric used in VQA challenge (Antol et al., 2015) is employed to evaluate the performance with the OK-VQA dataset. Furthermore, we evaluate the generated caption with BLEU (Papineni et al., 2002), CIDER (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004) metrics.

4.3 Uncertainty-based caption generation

	Corr
cap_sim* & al_un*	-0.1907
cap_sim* & ep_un*	-0.1653
al_un* & ep_un*	0.4518

*cap_sim represents a caption similarity. al_un and ep_un represent aleatoric uncertainty and epistemic uncertainty, respectively.

Table 2: Table of correlation with uncertainty and similarity

Table 3 shows performances of the baseline model for caption generation on OK-VQA dataset. When comparing the image caption performance on the OK-VQA dataset with Att2in, BuDn, and Transformer models, overall the Transformer model shows better performance than others. Our study uses Transformer model for uncertainty modeling. Fig 7 shows aleatoric uncertainty and epistemic uncertainty of the word of the generated caption, and the word of uncertain action and unusual object in the image shows higher uncertainty than the average uncertainty of the sentence.

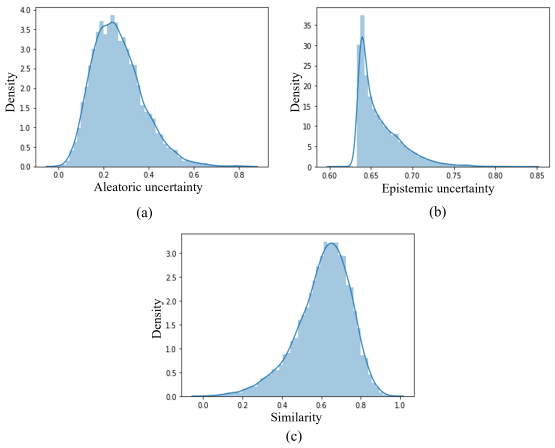

Figure 5: The distribution of the uncertainty of generated caption and similarity between the caption and the ground-truth caption

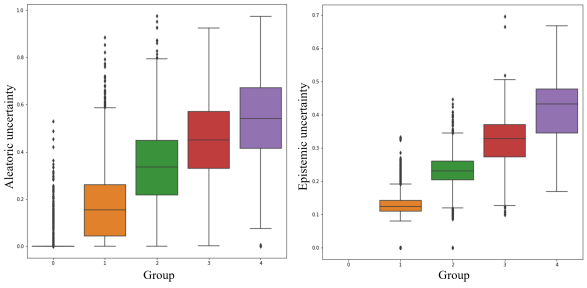

Figure 6: Boxplot of the uncertainty value according to the number of hallucinated objects

Table 2 illustrates correlation between uncertainty and caption similarity. The caption similarity and aleatoric uncertainty have a negative correlation of -0.1907, and the correlation between similarity and epistemic uncertainty is -0.1653. The correlation between aleatoric uncertainty and epistemic uncertainty shows a positive correlation, with a value of 0.4518. The correlation analysis shows the relationship between caption similarity and uncertainty. Using the caption similarity is also significant. In Fig 5, the distributions of (a) aleatoric uncertainty and (b) epistemic uncertainty show a right-skewed distribution, while (c) caption similarity distribution depicts a left-skewed distribution. Since there are extreme values in distributions, se-

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDER	METEOR	ROUGE-L
Att2in (Rennie et al., 2017)	0.7843±0.00005	0.6077±0.0002	0.4508±0.00032	0.3302±0.00038	1.0833+0.0016	0.2604±0.00018	0.5561±0.00019
BuDn (Anderson et al., 2018)	0.8123±0.00015	0.6516±0.00009	0.5017±0.00003	0.3786±0.00004	1.2527±0.00039	0.2858±0.00002	0.5859±0.00005
Transformer (Vaswani et al., 2017)	**0.8290±0.00028**	**0.6828±0.00036**	**0.5410±0.0004**	**0.4216±0.0004**	**1.3864±0.0012**	**0.2997±0.00013**	**0.6043±0.00023**

Table 3: Performances of image captioning with implicit commonsense knowledge on OK-VQA dataset

Caption: A fire **(0.40, 0.68)** hydrant on the side of a street,
m = (0.14, 0.64)

(a)

Caption: A white bird is standing on the top **(1.12, 0.82)** of an oven,
m = (0.36, 0.66)

(b)

Caption: A person **(0.812, 0.75)** is holding a teddy bear, m = (0.24, 0.67)

(c)

Figure 7: Image captioning results on OK-VQA dataset. a value in bracket is aleatoric uncertainty and epistemic uncertainty, respectively and m represents an average aleatoric uncertainty and an average epistemic uncertainty in a sentence, respectively

mantic inconsistency can be identified with uncertainties of the caption and the caption similarity.

This study analyzes the uncertainty relationship according to the number of hallucinated objects in the generated caption. The proportion of hallucinated objects of generated captions is calculated according to a synonym criteria of (Rohrbach et al., 2018). After synonym filtering of the generated caption, the number of hallucinated objects in the generated caption is counted. We divide the ratio of the number of words of the hallucinated objects among the caption words into 5 groups (0: 0∼0.2, 1: 0.2∼0.4, 2: 0.4∼0.6, 3: 0.6∼0.8, 4: 0.8∼1.0). We calculate the average uncertainty of the caption over the average uncertainty of the hallucinated objects. As shown in Fig 6, the more hallucinated objects in the caption, the higher aleatoric and epistemic uncertainty. We also perform qualitative analysis, as shown in Fig 7. For the example shown in Fig 7, the generated caption contains uncertain words with higher aleatoric and epistemic uncertainty than m (the average aleatoric uncertainty and the average epistemic uncertainty in a sentence).

4.4 KVQA with semantic inconsistency

We present our KVQA result with the proposed semantic inconsistency model. An ablation study is performed with three values of caption similarity, aleatoric uncertainty, and epistemic uncertainty with the weights in Eq. (7). In Table 4, the baseline model makes use of both explicit and implicit knowledge. The performance on the OK-VQA dataset shows 31.15% accuracy. When caption similarity is added, the accuracy increases by 0.4%. In addition, when aleatoric and epistemic uncertainty are added, respectively, it shows further improvement.

Also, when the similarity and aleatoric uncertainty are added, the accuracy increases by 0.49%.

Model	Accuracy
Baseline	31.15
Baseline + Cap_sim	31.55
Baseline + Aleatoric Uncertainty	31.28
Baseline + Epistemic Uncertainty	31.93
Baseline + Cap_sim + Epistemic Uncertainty (FC*)	31.64
Baseline + Cap_sim + Aleatoric Uncertainty (FC*)	**32.45**
Baseline + Cap_sim + Epistemic Uncertainty + Aleatoric Uncertainty (FC*)	31.07

*FC: Fully connected layer

Table 4: An external knowledge assimilation method ablation study on OK-VQA dataset

Q: Which item in this room is usually to wash hands?
A: sink
Baseline: bed
Ours: sink

(a)

Q: What is the purpose of these objects?
A: [decoration, art]
Baseline: tell time
Ours: decoration

(b)

Q: What is the purpose of the logos on this truck?
A: identification
Baseline: car
Ours: safety

(c)

Figure 8: Comparison with the predicted answers of the proposed method and the baseline model on OK-VQA dataset

Model	Accuracy
Q-Only	14.93
XNMNet (Shi et al., 2019)	20.67
BAN (Kim et al., 2018)	25.17
BAN + AN (Marino et al., 2019)	25.61
BAN + KG-Aug (Li et al., 2020)	26.71
MUTAN (Ben-Younes et al., 2017)	26.41
MUTAN + AN (Marino et al., 2019)	27.84
KA (Ziaeefard and Lécué, 2020)	29.03
ViLBERT (Lu et al., 2019)	31.35
KRISP* (Marino et al., 2021)	31.15
Ours	**32.45**

*Re-implemented result with the authors'code and parameter setting

Table 5: OK-VQA performance comparing with the state-of-the art approaches

The best performance is 32.45% accuracy when caption similarity and the epistemic uncertainty are concatenated. In addition, when the three values of caption similarity, aleatoric uncertainty, and epistemic uncertainty are added, the accuracy is 31.07%, which shows lower performance than that of the baseline. From the results, when all three values are given to a model, the model cannot predict a correct answer. As shown in Table 5, the model with both explicit knowledge, implicit knowledge, and semantic inconsistency method achieves the state-of-the-art performance. We also present a qualitative analysis of the model in Fig 8. We compare the prediction from our model with the baseline's. For (a) and (b), our model selects the correct answer. In addition, (c) for the proposed model predicted an answer that is more similar to the correct answer than the baseline model. Also, the proposed method predicts correct answers for (a) an image with a part of the sink, (b) an image with an unusual object illustrated in Fig 8.

5 Conclusion and future work

In this study, we propose a novel semantic inconsistency measure through uncertainty modeling and semantic similarity for KVQA that can make use of diverse KBs more effectively. As KBs are often incomplete or incompatible with the given problem, the use of knowledge should be moderated. With the proposed model, we achieve the state-of-the-art results on KVQA. In future work, we plan to further explore diverse ways of using KBs based on the characteristics of the KB and the given problem.

References

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 conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary" Ives. 2007. Dbpedia: A nucleus for a web of open data. In

The Semantic Web, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

Sumithra Bhakthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. 2020. Do dogs have whiskers? A new knowledge base of haspart relations. *CoRR*, abs/2006.07510.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Li Kai, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1184–1193. PMLR.

Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2020. Multi-modal graph neural network for joint reasoning on vision and scene text. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12743–12753.

Noa García and Yuta Nakashima. 2020. Knowledge-based video question answering with unsupervised scene descriptions. *ECCV*, abs/2007.08751.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5351–5359.

Drew A. Hudson and Christopher D. Manning. 2019. Learning by abstraction: The neural state machine. In *NeurIPS*.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11109–11116.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 5580–5590. Curran Associates Inc.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017a. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1227–1235.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

H. Liu and P. Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Álvaro García-Martín. 2020. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Raeid Saqur and Karthik Narasimhan. 2020. Multimodal graph networks for compositional generalization in visual question answering. In *Advances in Neural Information Processing Systems*, volume 33, pages 3070–3081. Curran Associates, Inc.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Jiaxin Shi, Hanwang Zhang, and Juanzi Li. 2019. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8376–8384.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. *ICLR*, abs/1908.08530.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

Liyang Zhang, Shuaicheng Liu, Donghao Liu, Pengpeng Zeng, Xiangpeng Li, Jingkuan Song, and Lianli Gao. 2021. Rich visual knowledge-based augmentation network for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4362–4373.

Maryam Ziaeefard and Freddy Lécué. 2020. Towards knowledge-augmented visual question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1863–1873.