QACE: Asking Questions to Evaluate an Image Caption

Hwanhee Lee[1], **Thomas Scialom**[2,3], **Seunghyun Yoon**[4], **Franck Dernoncourt**[4], **Kyomin Jung**[1]
[1]Dept. of Electrical and Computer Engineering, Seoul National University, Seoul, Korea
[2]Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[3]reciTAL, Paris, France, [4]Adobe Research, San Jose, CA, USA
{wanted1007, kjung}@snu.ac.kr, thomas@recital.ai
{syoon, franck.dernoncourt}@adobe.com

Abstract

In this paper, we propose QACE, a new metric based on **Q**uestion **A**nswering for **C**aption **E**valuation. QACE generates questions on the evaluated caption and checks its content by asking the questions on either the reference caption or the source image. We first develop QACE_{Ref} that compares the answers of the evaluated caption to its reference, and report competitive results with the state-of-the-art metrics. To go further, we propose QACE_{Img}, which asks the questions directly on the image, instead of reference. A Visual-QA system is necessary for QACE_{Img}. Unfortunately, the standard VQA models are framed as a classification among only a few thousand categories. Instead, we propose Visual-T5, an *abstractive* VQA system. The resulting metric, QACE_{Img} is multi-modal, reference-less, and explainable. Our experiments show that QACE_{Img} compares favorably w.r.t. other reference-less metrics. We will release the pre-trained models to compute QACE.[1]

1 Introduction

Image captioning is a task that aims to generate a description containing the main content of a given image. The field of caption generation is prolific (Vinyals et al., 2015; Anderson et al., 2018), and it is, therefore, important to provide reliable evaluation metrics to compare the systems. Most of the prior works still report n-gram similarity metrics such as BLEU (Papineni et al., 2002) or CIDEr (Vedantam et al., 2015). However, these n-gram similarity metrics often fail to capture the semantic errors in the generated captions (Novikova et al., 2017).

To overcome this limitation, we propose QACE, a radically different evaluation framework from n-gram metrics. QACE first generates questions about the candidate caption, and then checks if the answers are consistent w.r.t. either the reference or the source image. We depict QACE in Figure 1.

Specifically, we propose two variants of QACE, depending on what content the evaluated caption is compared to: QACE_{Ref} when it is compared to the reference, and QACE_{Img} when it is compared to the source image. QACE_{Img} has the desired feature to be *reference-less*, i.e., the score can be computed without requiring a gold reference.

In this reference-less setup, a Visual Question Answering (VQA) system is required to answer those questions. However, in the VQA literature (Antol et al., 2015), the task is usually seen as a classification task on 3k pre-defined answer choices (e.g., blue, sea, or banana). Therefore, these VQA models are not general QA systems; their usage off-the-shelf for QACE_{Img} would limit the comparison to these very few pre-defined categories, which is not satisfying. To solve this issue, we also propose an abstractive VQA system Visual-T5 as a new module for QACE_{Img} that can generate free-form abstractive answers given a textual question and an image. We conduct a human evaluation of Visual-T5 and show that it is capable of generating accurate abstractive answers. Using Visual-T5, we are now able to compare the answers of the candidate caption directly with the answers of the corresponding image.

Experimental results show that our proposed QACE_{Ref} and QACE_{Img} show promising results compared to other reference and reference-less metrics on three benchmark datasets: Pascal50s (Vedantam et al., 2015), Composite (Aditya et al., 2015) and Flickr8k (Hodosh et al., 2013). Also, as shown in Figure 1, QACE has a natural form of interpretability through the visualization of the questions and the answers.

2 Related Work

Image Captioning Metrics Similar to other text generation tasks such as machine translation and

[1]https://github.com/hwanheelee1993/QACE

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4631–4638
November 7–11, 2021.

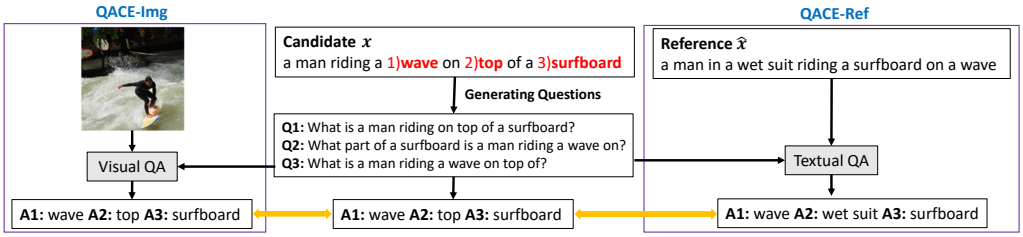


Figure 1: The overall flow of QACE. QACE extracts possible answer spans and generates answer-aware questions for a given candidate caption x. The VQA and TQA answer these questions given the image and reference captions, respectively. The correctness of the candidate caption is evaluated by comparing the answers.

summarization, n-gram similarity metrics such as BLEU, METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004) are arguably the standard in automatic evaluation. Among them, the most widely used metric is CIDEr (Vedantam et al., 2015) which uses TF-IDF based weighted n-gram similarity. SPICE (Anderson et al., 2016) metric is based on scene graph, while more recently, BERTScore (Zhang et al., 2019) compute the similarity of the contextualized embeddings. Different from prior works, we are the first to use Question Generation (QG) and Question Answering (QA) to evaluate the image captions.

Question and Answering for Evaluation Fisch et al. (2020) proposes a new method to generate informal captions that can answer the visual questions. In our work, we focus on caption evaluation using the QA systems, not on generating the captions. Several QA-based evaluation metrics (Scialom et al., 2019; Wang et al., 2020) are recently proposed to evaluate abstractive summarization. However, all those prior works are limited to text-to-text evaluation, while our work develops a multi-modal metric.

3 QACE

We propose QACE, which is a QG- and QA-based framework for evaluating an image caption. As shown in Figure 1, QACE first extracts answer candidates (i.e., 1) wave, 2) top, 3) surfboard) from a candidate caption and generates corresponding questions. With these questions, visual-QA (VQA) and textual-QA (TQA) models answers given their context (i.e., image and reference $\hat{x}$). By comparing the answers from each source, we can directly judge the correctness of the candidate caption.

3.1 Question Generation

The goal of this component is to generate questions that ask the primary information of the candidate caption. Our QG model is a text-to-text generation model (i.e., T5 (Raffel et al., 2020)), fine-tuned on SQuAD v2 (Rajpurkar et al., 2018) to generate answer-aware questions. Given a caption, we extract possible answer span; in particular, we focus on extracting noun phrases since they mostly contain salient information and can be easily foiled (Shekhar et al., 2017). We argue that questions generated on this salient information should be answered similarly from the image or the captions if they share the same information.

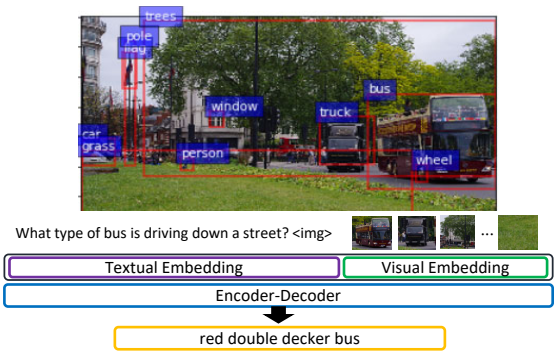


Figure 2: The overview of Visual-T5, an abstractive VQA model. We embed questions with additional special separation token and concatenate the visual embeddings to make inputs for T5.

3.2 Question Answering

For QACE_{Ref}, we use a TQA model. We train T5 to answer the generated questions (see 3.1) with the reference captions as context. Conversely, QACE_{Img} requires a VQA model. We propose a new architecture, Visual-T5, that can generate abstractive answers given an image and a question, as opposed to the standard multiple-choice VQA.

3.3 Abstractive Visual Question Answering

When no reference captions are available, one of the most important parts of QACE is the VQA model that can produce correct answers. To move beyond VQA as a classification task, we are the

first, to the best of our knowledge, to develop an abstractive VQA model that can generate free-form answers. Specifically, we enable multimodal encoding for T5, inspired by the previous works on adapting pre-trained language models for multimodal tasks (Scialom et al., 2020). We illustrate our proposed Visual-T5 in Figure 2. Based on default T5 architecture, Visual-T5 has an additional visual embedding layer that encodes regional features of the image from Faster RCNN (Ren et al., 2015). This linear layer maps detection features to 768 dimensions, same as the dimension of textual embedding. This 768d features are therefore considered as a standard token in Visual-T5, which can encode an image and a question together. We provide more details in Appendix.

3.4 QACE Metric

For a given candidate caption x, We use QG to generate questions $Q = (q_1, ..., q_M)$ for all of M noun phrases of x. Then, we compare the answers for each question in Q on x with the answers on the reference source. We introduce two QACE variants, QACE_{Ref} for which the reference caption is compared, and QACE_{Img} for which the source image is compared. Using QG and QA, we compute QACE_{Ref} and QACE_{Img} as follows:

$$QACE = \frac{\Sigma_{i=1}^{M} f(QA(q_i, x), QA(q_i, ctx))}{M}, \quad (1)$$

where ctx corresponds to the image for QACE_{Img} and the gold reference for QACE_{Ref}, $f(A_1, A_2)$ is the function that measures the similarity between two answers A_1 and A_2. The standard metric in QA is the F1, as introduced by Rajpurkar et al. (2016). However, two abstractive answers can be similar but written in two different ways, limiting the effectiveness of a naive F1. Hence, in addition to the F1, we propose to use the BERTScore. Finally, we also complete the similarity metrics using the answerability of the questions for function f, in order to measure whether the question is answerable. The answerability corresponds to $1 - P_{unanswerable}$, where $P_{unanswerable}$ is the probability attributed by the model to the token *unanswerable*.[2] To consider all the different aspects, we use the average of three values computed using each function as the default value of QACE.

[2]SQuAD v2 contains unanswerable questions, for which we associate the token *unanswerable* as the correct answer during training. Therefore, our QA model associates this token with the probability that the question is not answerable.

4 Synthetic Data Generation for VQA

As discussed in 3.3, relying on a VQA dataset such as VQA v2 (Goyal et al., 2017) limits possible answers to a small size of pre-defined categories. To train a general and abstractive VQA model, we create synthetic abstractive VQA datasets. We generate Questions/Answers pairs using the captions in the training set of MS-COCO (Lin et al., 2014). Specifically, we extract noun phrases from a reference caption and generate an answer-aware question using our QG model. To increase the validity of these synthetic questions, we apply the round trip consistency (Alberti et al., 2019), filtering out the questions for which the QA model predicts a different answer than the extracted noun phrase. We convert these synthetic QA dataset to create {*question, answer, image*} triples by concatenating the corresponding images to these captions.

In addition, we randomly add 20% of unanswerable questions[3] to the synthetic training set, so that the model learns to judge the answerability of a given question. Through this, if a candidate caption contains any hallucinating content that is not included in the image, questions about it can be marked as unanswerable by our VQA model, as shown in the second example of Figure 3. This synthetic dataset enables the training of the abstractive VQA model. We report the performance of the model through a human evaluation in Section 5.2.

5 Experiments

5.1 Benchmark Dataset

We evaluate our proposed metric on three benchmark datasets (i.e. human annotations), *PASCAL-50S*, *Composite* and *Flickr8k*.

PASCAL-50S provides 4k caption triplet <*A*, *B*, *C*>, where "*A*" is composed of 50 reference captions(*A*) and two candidate captions(*B*, *C*) for the given image. There are human judgments as to which "*B*" or "*C*" is more appropriate caption for a given image compared to "*A*".

Composite is composed of 11,985 human judgments scores range from 1 to 5 depending on the relevance between each candidate caption-image pair with 5 reference captions.

Flickr8k provides three human-expert judgments for 5,822 candidate caption-image pairs. The scores are from 1 to 4, depending on the relevance of each

[3]We consider an image and a question that are not paired to be unanswerable, and do negative sampling.

	Ref?	Pascal50s	Composite	Flickr8k
BLEU-4	✓	65.2	45.7	28.6
ROUGE-L	✓	67.7	47.7	30.0
METEOR	✓	**80.5**	46.6	40.3
CIDEr	✓	77.8	47.4	41.9
SPICE	✓	76.1	48.6	**45.7**
BERTScore	✓	72.0	45.6	30.5
QACE-Ref (ours)	✓	75.1	49.3	40.5
F1	✓	57.5	**55.1**	9.2
BERTScore	✓	76.4	46.0	30.9
Answerability	✓	71.6	47.3	39.0
-Perplexity	✗	46.8	1.7*	10.1
VIFIDEL	✗	69.0	13.1	**33.6**
QACE-Img (ours)	✗	**70.0**	**19.1**	29.1
F1	✗	62.0	12.5	27.3
BERTScore	✗	65.9	12.8	27.1
Answerability	✗	**74.5**	15.7	27.8

Table 1: First column represents the accuracy of matches between human judgments in PASCAL50s. Columns 2 to 3 show the Kendall Correlation between human judgments and various metrics. All p-values in the results are < 0.05 except for *.

caption-image pair.

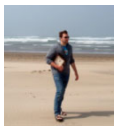
5.2 Results and Discussions

Computation Details For all of the results on reference based metrics we reported in the paper, we compute the average of each metric score with each reference for all of the references on each dataset.

QACE Performance We compare our proposed method with the following widely used metrics: BLEU-4, ROUGE-L , METEOR, CIDEr, SPICE, and the BERTScore. We present the experimental results for all three datasets in Table 1. For the reference-aware metrics, QACE_{Ref} shows best results on Composite and comparable to the best metrics for Pascal50s and Flickr8k, indicating the relevance of a QA based metric to evaluate image captioning.

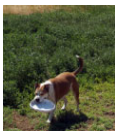
For the reference-less metrics, all the correlations are lower this time, showing the difficulty of evaluating the captions without reference. Nonetheless, among these metrics, QACE_{Img} shows the best results for Pascal50s and Composite and comparable results in Flickr8k. For Flickr8k, we found that more than half of the human judgments of the candidate captions are less than 0.2 as 0 to 1 scale. In other words, most of the captions in this dataset are totally not related to the image. For this reason, most of the generated questions are unanswerable for an image and we explain that this leads to relatively lower performance of QACE_{Img} in Flickr8k compared to other metrics.

Furthermore, We investigate the independent contribution of each answer similarity function, f,



Candidate: a man^{A1} is standing on a sunny beach^{A2} **(Human: 1.0)**
Reference: a man walks down the beach near the ocean
Q1: What is standing on a sunny beach?
Q2: What is a man standing on?

Ref	**A1**:man **A2**:beach	$QAEC_{Ref}$: 0.88
Img	**A1**:man **A2**:sand	$QAEC_{Img}$: 0.79



Candidate: a cow^{A1} is standing in a field^{A2} of grass^{A3} **(Human: 0.2)**
Reference: a dog with a frisbee standing in the grass
Q1: What animal is standing in a field of grass?
Q2: What is a cow standing in?
Q3: What type of field is a cow standing in?

Ref	**A1**:dog **A2**:grass **A3**:grass	$QAEC_{Ref}$: 0.60
Img	**A1**:dog **A2**:unanswerable **A3**:grassy field	$QAEC_{Img}$: 0.47

Figure 3: Case study on QACE metric. Human judgments are normalized to between 0 and 1.

in computing QACE and present the results in Table 1 (note that default QACE-Img uses the mean of F1, BERTScore and answerability). The table reveals that each similarity function has a different aspect, and averaging three results suggests the best performance for two of three datasets.

VQA Model Performance Visual-T5 is one of the main components of QACE_{Img}. Since it can generate free-form answers, its automatic evaluation is challenging. We therefore conduct a human evaluation on 200 examples randomly sampled from the test set. We hire three annotators to judge whether the generated answer is correct or not given the image. On the majority vote from three annotators, VQA model correctly answers for the average 69% of the examples. Among these 69% correct answers, half of them were written differently from the original answer, showing that our model can generate abstractive answers.

Case study Different from the previous metrics, QACE can be easily interpreted through the visualization of the generated questions and the following answers as shown in Figure 3. In the first example, we observe that the second question is answered differently by the VQA model (*sand* VS *beach*). Despite, the answer itself being correct - it is true that the man is standing on the sand - it results in a lower score for QACE_{Img} compared to QACE_{Ref}. This emphasizes the importance to use other similarity metrics than the F1 when comparing two answers (see Section 3.4). For instance, BERTScore should be able to consider closer *sand* and *beach* than *sand* and a random word.

The second example is very illustrative: for the first question, both TQA and VQA answer *dog*, hence detecting an error in the candidate caption that talks about a *cow*. The second question refers to the *cow*, which makes it ambiguous. The VQA model considers it as *unanswerable*, while the TQA model correctly answers *grass*. Following

this study, we expect that $QACE_{Img}$ can be improved through a finer answer comparison method in future work.

6 Conclusion

In this paper, we propose QACE, a captioning metric that directly compares each content in the candidate caption with either the source image or a gold reference caption by asking questions. To enable asking questions directly on the source image, we introduce Visual T5, an abstractive VQA model to generate free-form visual answers, for which we report strong results based on a human evaluation. Our proposed metric can be applied in both reference and reference-less settings. It holds high explainability and compares favorably to the state-of-the-art in terms of correlations with human judgments.

Acknowledgements

We thank anonymous reviewers for their constructive and meaningful comments. K. Jung is with ASRI, Seoul National University, Korea. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1A2C2008855). This work was partially funded by gifts from Adobe Research.

Ethical Considerations

We compensate the workers with competitive pay, which is above hourly USD $10 for the human evaluation of VQA model. Furthermore, we used public datasets to train the models.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic qa corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

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.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan H Clark, and Regina Barzilay. 2020. Capwap: Captioning with a purpose. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8755–8768.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 91–99.

Thomas Scialom, Patrick Bordes, Paul-Alexis Dray, Jacopo Staiano, and Patrick Gallinari. 2020. What bert sees: Cross-modal transfer for visual question generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 327–337.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3237–3247.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265.

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.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Experimental Details

A.1 Reproducibility Checklist

Source Code We attach the source for computing QACE and training Visual-T5. In the question generation components, we use the Noun Chunks extractor from spaCy.[4]

Computing Infrastructure We use AMD Ryzen Threadripper 2950X (3.50 GHz) with GeForce RTX 2080 Ti for the experiments. The software environments are Python 3.6.6 and PyTorch 1.3.1.

Average runtime for each approach It takes average one second to generate all questions for a given candidate caption using a pre-trained question generation model. And it takes average about 0.1 seconds to compute visual and textual answers, and comparing the answers. For training VQA model, Visual-T5, each epoch takes 40 minutes using a single RTX 2080 Ti GPU.

Hyperparameters We use the pre-trained *t5-base* for question generation and TQA model in the model repository[5] of huggingface (Wolf et al., 2020). We use *t5-small* to fine-tune our VQA model. Based on *t5-small*, we added single linear layer to encode visual features and then train the model for 5 epochs with batch size of 128. The number of the synthetic training set is 1 million and we split the dataset into 9:1 proportion for training and validation. For the max sequence length, we set 64 to the input sequence including the visual tokens, and set 32 to output sequence.

Number of Model Parameters The number of parameters for QG is 222.9M, TQA is 222.9M and VQA is 61.6M.

A.2 Significance Test

We conduct a standard way to test the significance of the correlation coefficient for all of the reported correlation coefficients in the paper. We use a t-test that uses a null hypothesis, which is an absence of association, and report the p-value for each coefficient.

B Abstractive Visual Question Answering

We provide the training details including the additional output examples of our proposed abstractive VQA model, Visual-T5 in this section.

[4]https://spacy.io/usage/linguistic-features#noun-chunks
[5]https://github.com/mrm8488/question_generation

Figure 4: Various output examples on the evaluation set of abstractive VQA model, Visual-T5.

B.1 Visual Embedding

We extract the regional features for each object using Faster RCNN (Ren et al., 2015). We fixed the number of boxes to 36 and each regional feature consists of dimension 2048 and 6 additional dimensions consists of the location and the size of each box. We concatenate this additional dimensions to make dimension of 2054 for each regional feature. And single linear layer maps these 2054d features to 768d to be considered as a token in T5.

B.2 Answer Examples

We provide more examples of our abstractive VQA models in Figure 4. We observe that many predicted answers are correct, but expressed in a different form as in the first and the second example. Also, model outputs *unanswerable* to the questions that are unanswerable for a given image like the third example.

B.3 Answerability

We make unanswerable visual questions by randomly sampling the questions from the different images to the given image. We mixed 20% of these unanswerable questions similar to the third example in Figure 4 to train VQA model.

B.4 Human Evaluation

We hire the workers whose locations in one of the US, UK, CA, NZ, AU to guarantee the fluency

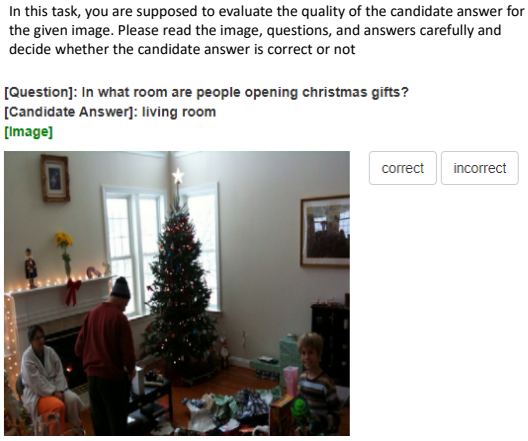


Figure 5: Full instructions and interface to workers for evaluating the answers of VQA model.

in English. We restrict the workers whose HIT approval rates are higher than 95%, and minimum hits are over 500. We pay workers more than USD $10 in an hour through several preliminary experiments on the compensation. We provide the full instructions and the interface in Figure 5. We compute the annotator agreement using Krippendorff's α (Krippendorff, 1970). We observe that Krippendorff's α is 0.56 that indicates a "moderate" agreement according to one of the referenced guidelines (Landis and Koch, 1977) for kappa-like measures.