

ICE: A Benchmark for Human-like Image Commenting

Qiuyuan Huang¹, Yandong Guo², Oussama Elachqar³, Kedhar Nath Narahari⁴, Puneet Agrawal⁴, Jianfeng Wang², Hui Zhang⁵, Sonam Damani⁴, Kazushige Ito⁶, Donald Brinkman³, Xiaodong He⁷, Lei Zhang², Yu-Ting Kuo³

¹Microsoft Research AI; ²Microsoft Research; ³Microsoft, USA; ⁴Microsoft, India; ⁵Microsoft, China; ⁶Microsoft, Japan; ⁷JD AI Research
qihua@microsoft.com; yandong.guo@microsoft.com; ouelachq@microsoft.com;
kedharn@microsoft.com; punagr@microsoft.com; jianfw@microsoft.com;
huizh@microsoft.com; sodamani@microsoft.com; Kazushige.Ito@microsoft.com;
donaldb@microsoft.com; xiaodong.he@jd.com; leizhang@microsoft.com;
yuting.kuo@microsoft.com

Abstract. With the growing usage of images in conversations with Artificial intelligent(AI) agents, it has become imperative for AI agents to respond to images in more human-like manner. To achieve this, agents should not only describe the facts of an image but also express emotions and opinions about it. Despite phenomenal progress in the past several years, most existing image captioning systems mainly describe the concepts, objects or actions present in an image. Hence, they have an inherent drawback of neither being human-like nor facilitating deeper engagement. To overcome this limitation, we introduce a novel task of Image CommEnting (ICE) with the objective of generating a human-like comment given an image. To further facilitate research in this domain, we release a dataset, ICE v1, which includes one million images and their corresponding real human comments. We also explore how different existing image commenting techniques perform with this dataset and establish strong baselines for this task.

Keywords: Image commenting, image caption, LSTM, reinforcement learning, Tensor Product Representation (TPR), social intelligence

1 Introduction

Great progress has been made in the interdisciplinary research at the intersects of computer vision and natural language processing, especially image captioning, which is to produce literal descriptions for images. Good performance of image captioning has been achieved in [1–8], with the help of language modeling, deep neural network, and large datasets. However, we argue that image captioning may not be the natural response of human when a visual signal is given.

Image captioning focuses on literal descriptions of images. In the COCO caption dataset [9, 10], which is the de facto standard dataset for image captioning,

the ground truth has been generated by requesting human labelers on Amazon MTurk to “describe **all** the important parts of the scene”, and the same guidance is used to evaluate the machine-generated caption as well. Therefore, to visually detect all the facts, such as objects, actions, etc., in the image and design reasonable language models is critical for image captioning. For example, a good caption for the image in Fig. 1. is “some houses are on the foot of mountains”. This is of great value in helping people who have a low vision or can not see in certain environments, but may not be what a person would naturally respond to the image. For example, when a human is given the image in Fig. 1, a natural response might be a short comment like “gorgerous”, “nice picture”, or a question “where is this picture taken?”.



Fig. 1. Examples illustrating difference in Image Captioning and Image Commenting

To move beyond the literal description of the image content, we introduce a novel task of Image CommEnting (ICE), where given an image, the objective of the computer is to give a comment which a person would naturally provide given this image. While these comments are based on visually detecting and recognizing the facts of the image, they are beyond fact detection and recognition. Especially, the comments have rich styles, often express opinions, sentiments and emotions, and also serve the purpose of increasing user engagement.

The study of image commenting is of great value. First, in order to generate comment which person would naturally provide, an image commenting system needs to not only recognize the facts in the image, but also inference what a person would feel based on the factual recognition from the vision signal. There has been some early effort in this direction, for example, moving from image quality metric to aesthetic assessment, generating relevant questions based on images to express curiosity. However, image commenting provides a unique challenge to comprehensively understand the feeling of human given visual stimuli.

Second, an image commenting system needs to have certain level of interpersonal skill, be aware of common sense, and demonstrate a social personality. Moreover, image commenting also poses a great challenge to the research com-

munity due to noisy data, unreliable labeling, and large variations in subjective evaluation. All the above challenges are of great research value. Image commenting task may imbue the computer system to actually understand and think with sentiments, which is a hallmark of human cognitive system. We hope that the image commenting task could defend the Turing test from being criticized that “cannot prove that a machine can think” [11]. The image commenting system also has great value for many practical applications to improve user engagement, especially in social media and conversational agents, e.g., Microsoft commercial chat-bot and Facebook Messenger.

To further facilitate the image commenting research, we provide a training dataset¹. In this dataset, there are about one million images and all the images are associated with real comments generated by human in a natural way. The comments in the dataset have very different properties compared with captions in the COCO dataset, which demonstrates the necessity of our task from another perspective. Details are provided in the section 3. The same training data has been used to train models which has been used in one of the most popular commercial chat-bot and improved the number of conversation-turns per session and the accuracy of answers, which proves the value of the dataset to be published.

In this paper, we also introduce some baseline methods used to train an image commenting system with our training dataset. Moreover, we provide a test dataset and propose metrics to evaluate such a system. The preliminary experimental results and analysis are listed in the experiments section.

Our contributions in this paper are summarized as follows.

- We define the novel image commenting task.
- We provide both the training and testing datasets to facilitate the research in the area. The same training data has been used to train model in commercial systems.
- We list some baseline methods and evaluation metrics for the image commenting task.

The rest of the paper is organized as follows. In Section 2, we review the related work. Section 3 presents the method we used to collect the dataset and the properties of our dataset. We describe several baseline methods in Section 4 and list the corresponding experimental results in Section 5. Finally Section 6 concludes the paper with more discussions for the future work.

2 Related Work

Several tasks have been defined in the domain of the image(video)-to-text generation. Probably the most established task is that of the image captioning, of which the task is to generate a sentence/caption to literally describe a given image. As we discussed in the introduction, the caption generation focuses on

¹ The link for downloading the dataset will be published upon acceptance of the paper.

describing the (visual) facts in the image, rather than the natural responses of human. Detecting and recognizing objects and events are very critical for the task [9]. This might be the reason that most of the existing deep learning-based image captioning methods [1–8] is started with a visual feature extractor or word detector, then followed by a language model. Another task which is also focused on factual description is visual storytelling, where the task is to generate a sequence of coherent sentences (*i.e.*, a story) for an ordered image stream [12–14]. In this task, recurrent neural network (RNN) is usually deployed to leverage the temporal information [15, 16, 13].

The SentiCap in [17] and StyleNet in [18] are some early efforts to extend the factual description with sentiments or styles. In [17], Mathews et al. proposed to add positive and negative sentiments to the captions. Their training data is crowd-sourced on Amazon Mechanical Turk (AMT), where workers are requested to incorporate one or more (emotions/sentiments) into any one of the five existing captions. In [18], Gan et al. prepare the training data by requesting annotators to revise the **factual** captions to make it romantic or humorous. Therefore, we argue that there above early efforts essentially focus on providing factual descriptions, though extend the descriptions to be positive or negative, or to be romantic or humorous.

The Visual Question Generation (VQG) task might be more correlated with our task. The VQG system is intended to generate a natural and engaging question for a given image [19]. Though the VQG task also looks for the natural response and further engagement, which is similar to our Image CommEnting (ICE) task, the VQG task target only generating “questions”. As described in [19], most of the sentences (questions) in the VQG datasets start with the words “what”, “is”, “how”, “where”, and “why”. On the contrary, our Image CommEnting (ICE) task is more general as compared to VQG, since question is only one type of comment people would naturally provide. We list more details in the dataset property section. Note that VQG here is different from the general Visual Question Answering (VQA) in [20]. VQA is a task where the system is trained to provide an answer to a question about the image content. The question (pre-annotated by human) has to be visually verifiable. Another related work is COCO-QA (CQA) [21]. CQA does not use human-authored questions, but generates questions automatically from image captions of the COCO dataset by applying a set of transformation rules to generate the wh-questions.

3 Benchmark Task

We define our image commenting (ICE) task as to generate natural comments for a given image. The comments express opinions, sentiments, and emotions, and are intended to encourage the further conversational engagement. In order to facilitate the research for the defined task, we have prepared a dataset of about one million image-comment pairs and split them into training, validation, and testing. We also discuss metric options to evaluate the quality of the image commenting system.

3.1 Dataset Preparation

Many image-text generation datasets rely on hiring annotators to label under certain guideline [9, 17–19]. However, since we need to simulate the natural comments from human, we do not choose the controlled environment to construct our dataset.

On the contrary, we built our dataset by mining the information from the internet. More specifically, we scraped more than one million image-comment pairs from Reddit². Afterwards, we apply filters to both the images and comments to minimize the chance of containing sensitive content such as individuals name, adult information, politically sensitive, or ethnic-religious content, or other potentially offensive or contentious material, such as inappropriate references to violence, crime and illegal substances. After filtering, our dataset has roughly one million pairs which are then normalized for special characters, abbreviations etc. using NLTK [22], with the steps summarized as follows.

1. Change a verb to its simple present tense. Do not change a present participle or past participle form to its simple present tense.
2. Change each capitalized letter to its lower case.
3. Change an abbreviation to its non-abbreviated form. E.g., change “it’s” to “it is”.
4. Change multiple consecutive period marks to one period mark.
5. Remove special characters such as “\$”, “%” and “&”.

3.2 Dataset Properties

In this subsection, we analyze properties of our Image CommEnting (ICE) dataset to highlight the uniqueness of our task.

Sentiment word First, we list the top 40 words for both the comments from our ICE data and the captions from the COCO dataset [9] to intuitively understand the different between the datasets.

As shown in Fig. 2, there is almost no overlap among the top 40 words from the two sets. The most popular words in the comments in our ICE dataset is “like”, which is a word with strong positive sentiment. Referring to the sentiment word list from [23], we find 11 words among the top 40 words are sentiment words, as shown in Table 1. Interestingly, all the 11 words express positive sentiment. This also reveals a bias in the real scenario: which is that the usual comments tend to be of positive sentiment or that people are likely to have a positive attitude in conversations. On the contrary, the most frequent words in the COCO data tends to be the words used to describe facts, such as event facts including “sitting”, “holding”, etc., or objects, e.g., “people”, “woman”, etc. We apply the same sentiment word table in [23] and find that the only sentiment word is “top”. However, we argue that the word “top” in the caption dataset is often used to describe in the context of statements like “A is on top of B”, rather the positive opinion such as “top performers”.

² <https://www.reddit.com/>

Fig. 3(b) shows that the comments in the ICE dataset has the largest vocabulary size. This is expected due to the large size of our dataset (roughly one million comments), and also due to the fact that comments tend to be more diverse.

Moreover, Fig. 3(c) shows that the verb part of speech (POS) is represented with high percentage in the ICE dataset; this may be because verbs can help facilitate human conversation. Fig. 3(d) indicates that the ICE dataset uses significantly less abstract terms than other datasets; this may be because the sentences in the ICE dataset are more like oral language, to engage human social interaction.

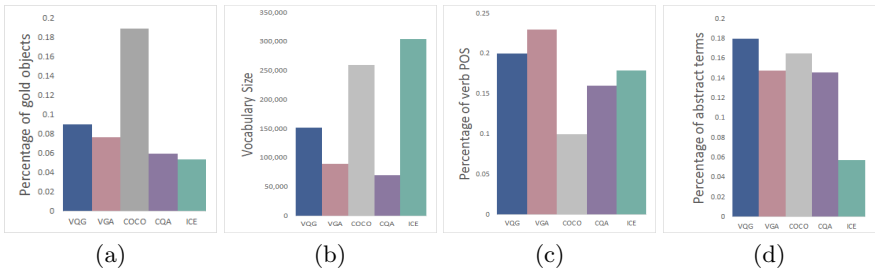


Fig. 3. Comparison of various annotations on the ICE dataset: (a) percentage of gold objects used in annotations, (b) vocabulary size, (c) percentage of verb POS, (d) percentage of abstract terms

Length We also calculate the length of both the comments in ICE and captions in COCO. Fig. 4 shows the histogram of the number of tokens of the comments in the ICE dataset, and the number of tokens of the captions in the COCO dataset. We found that on average, a comment has 12.22 tokens, while a caption in COCO has 10.45 tokens on average. Moreover, the comments in ICE have large variation in terms of length, compared with the captions in COCO. These facts may be attributed to the fact that COCO was constructed in a controlled environment with clear guidelines, while the ICE data represents the natural human behavior in a free environment.

From the above analysis, it is evident that the ICE dataset, though focused on image-to-text generation, has very different properties than other datasets in terms of special focus on sentiment expression, preference on social engagement, and larger variation in terms of topics. We hope that the release of this dataset would facilitate research along this direction.

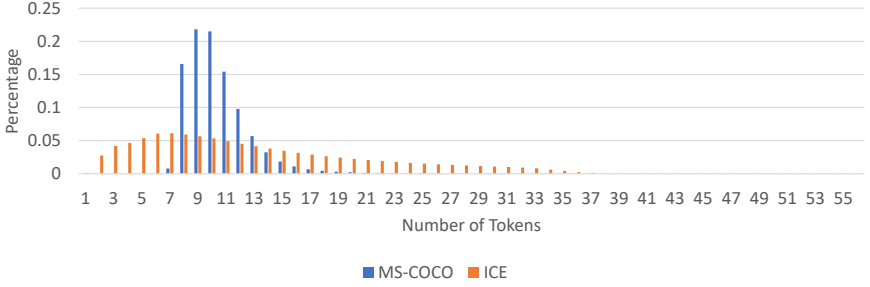


Fig. 4. Histogram of the length of the comments in ICE data set and the length of the captions in COCO data set

4 Baseline Methods

In order to validate our benchmark task and the associated dataset, we apply three baseline methods to train image commenting model. Two methods are using the Long-Short-Term Memory (LSTM), while the other one is using the Tensor Product Generation Network (TPGN). Here we briefly revisit these baseline methods for readers’ convenience and the completeness of the paper.

We denote the input image with $\mathbf{I}$ and the associated comment with $\mathbf{X}$. Typically, the image is transferred into a vector with a deep convolutional neural network (CNN), while the comment $\mathbf{X}$ is transferred with a vocabulary followed by another transfer, as shown in Fig. 5. Assume that comment $\mathbf{X}$ consists of T words. We define $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T]$, where $\mathbf{x}_t$ is a one-hot encoding vector of dimension V and V is the size of the vocabulary. The length T may be different for different captions. The t -th word in a comment, $\mathbf{x}_t$, is embedded into an n_x -dimensional real-valued vector $\mathbf{w}_t = \mathbf{W}_e \mathbf{x}_t$, where $\mathbf{W}_e \in \mathbb{R}^{n_x \times V}$ is a word embedding matrix.

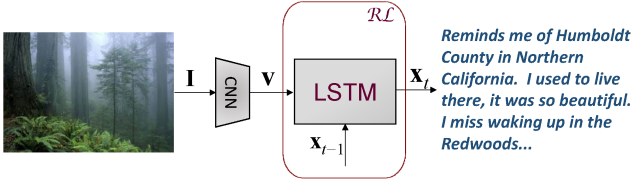


Fig. 5. Architecture of a baseline LSTM

An LSTM model can be trained through minimizing the cross-entropy (XE) loss function [24–26]. We call this method as LSTM-XE. Another option is to

directly optimize for task specific metrics such as CIDEr [27] and address the exposure bias problem, Reinforcement Learning (RL) can be used to train an LSTM model [28]. At each time step, the word generated by the model can be viewed as an action taken by an RL agent. Once the full sentence is $\hat{\mathbf{X}}$ is generated, it is compared against the ground truth $\mathbf{X}$. Under the RL terminology, the LSTM is regarded as an “agent” that interacts with an external “environment” (e.g., words and image features). The parameters of the LSTM, θ , define a policy p_θ , that produces an “action” (e.g., prediction of the current word). After each action, the agent updates its internal “state” (e.g., cell and hidden states of the LSTM). Upon generating the end-of-sentence (EOS) symbol, the agent receives a “reward” (e.g., the CIDEr metric of the generated sentence w.r.t. the ground-truth sentences); denote this reward by r . The objective of RL-based training is to minimize the negative expected reward:

$$L(\theta) = -\mathbb{E}_{\mathbf{X}^s \sim p_\theta} [r(\mathbf{X}^s)] \quad (1)$$

where $\mathbf{X}^s = [\mathbf{x}_1^s, \dots, \mathbf{x}_T^s]$ and $\mathbf{x}_t^s$ is the word sampled from the distribution p_θ at time t . In order to compute the gradient $\nabla_\theta L(\theta)$, one can use the REINFORCE algorithm [29] as below

$$\nabla_\theta L(\theta) = -\mathbb{E}_{\mathbf{X}^s \sim p_\theta} [r(\mathbf{X}^s) \nabla_\theta \log p_\theta(\mathbf{X}^s)] \quad (2)$$

To reduce the variance of the gradient estimate, a reference reward or baseline b is introduced to the gradient as below:

$$\nabla_\theta L(\theta) = -\mathbb{E}_{\mathbf{X}^s \sim p_\theta} [(r(\mathbf{X}^s) - b) \nabla_\theta \log p_\theta(\mathbf{X}^s)] \quad (3)$$

In Ref. [28], the authors proposed a self-critical sequence training (SCST) approach, which has $b = r(\hat{\mathbf{X}})$ where $r(\hat{\mathbf{X}})$ is the reward obtained by the current model under the inference algorithm used at test time. Then (3) becomes

$$\nabla_\theta L(\theta) = -\mathbb{E}_{\mathbf{X}^s \sim p_\theta} [(r(\mathbf{X}^s) - r(\hat{\mathbf{X}})) \nabla_\theta \log p_\theta(\mathbf{X}^s)] \quad (4)$$

In practice, the expected gradient can be approximated by a single Monte-Carlo sample $\mathbf{X}^s$ from p_θ as below

$$\nabla_\theta L(\theta) \approx -(r(\mathbf{X}^s) - r(\hat{\mathbf{X}})) \nabla_\theta \log p_\theta(\mathbf{X}^s) \quad (5)$$

Minimizing the reinforcement loss (or maximizing rewards) does not ensure the readability and fluency of the generated caption [30]. Using a mixed loss function, which is a weighted combination of the cross-entropy (XE) loss L_{XE} and the reinforcement learning (RL) loss L_{RL} , helps improve readability and fluency since the cross-entropy loss is based on a conditioned language model, with a goal of producing fluent captions. That is, the mixed loss function is given by

$$L(\theta) = \gamma L_{RL}(\theta) + (1 - \gamma) L_{XE}(\theta) \quad (6)$$

where $\gamma \in [0, 1]$. For annealing and faster convergence, we start with $\gamma = 0$ (i.e., minimizing the cross-entropy loss), and then gradually increase γ to a maximum value $\gamma_{max} < 1$. We call the LSTM trained by the aforementioned reinforcement learning with the mixed loss function in (6) as LSTM-RL.

Fig. 6 shows the architecture of the third baseline scheme – TPGN [31], which is based on Tensor Product Representations (TPR) [32, 33]. TPR is a structured neural-symbolic model developed in cognitive science over 20 years ago. In the TPR theory, a sentence can be considered as a sequences of *roles* (i.e., grammatical components) with each filled with a *filler* (i.e., tokens). Given each role associated with a pair-wisely orthogonal *role vector* $\mathbf{r}_t$ and each filler associated with a *filler vector* $\mathbf{f}_t$, the TPR of a sentence can be computed as $\mathbf{S} = \sum_t \mathbf{f}_t \mathbf{r}_t^\top$. Comparing with the popular RNN-based representations of a sentence, a good property of TPR is that decoding a token of a timestamp t can be computed directly by providing an *unbinding vector* $\mathbf{u}_t = \mathbf{r}_t / \mathbf{r}_t^\top \mathbf{r}_t$. That is, $\mathbf{f}_t = \mathbf{S} \cdot \mathbf{u}_t$. Under the TPR theory, encoding and decoding a sentence is equivalent to learning the role vectors $\mathbf{r}_t$ or unbinding vectors $\mathbf{u}_t$ at each position t .

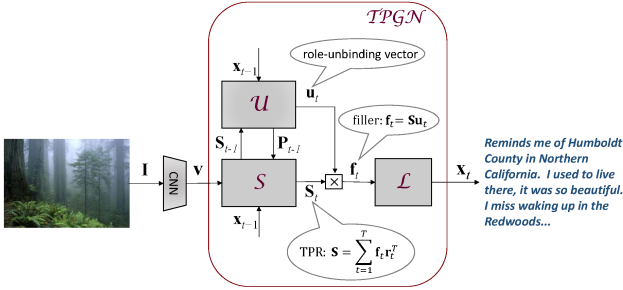


Fig. 6. Architecture of TPGN. “ $\otimes$ ” denotes the matrix-vector product. $\mathcal{S}$ is a sentence-encoding subnetwork, which could host a TPR of the sentence to be generated. $\mathcal{U}$ is an unbinding subnetwork, which could output a sequence of unbinding vectors $\mathbf{u}_t$. $\mathcal{L}$ is a lexical-decoding subnetwork, which converts the embedding vector $\mathbf{f}_t$ to the 1-hot vector $\mathbf{x}_t$ corresponding to the word x_t

5 Experimental Results

5.1 Evaluation

For our experiment, we split our ICE dataset into training (about one million image-comment pairs), validation (five thousand) and testing (five thousand). The data split will be released along with the dataset. We use a vocabulary that consists of 18,018 words.

During the experiment, we used ResNet-152 [34], pretrained on the ImageNet dataset, to extract image features. The feature vector $\mathbf{v}$ has 2048 dimensions.

Word embedding vectors are obtained by Google word2vec tool [35]. The model is implemented in TensorFlow [36] with the default settings for random initialization and optimization by back-propagation.

For the image captioning method, we apply the same method (LSTM-XE) to the COCO dataset, and test with our dataset just to validate the difference between the two datasets.

Table 2. Performance of the baseline models on the ICE dataset

Methods	BLEU-4	ROUGE-L	CIDEr-D	METEOR-1	METEOR-2	SPICE
Image captioning	0.017	0.11	0.0111	0.0432	0.0957	0.0293
LSTM-XE	0.0312	0.132	0.0184	0.0527	0.1109	0.0305
LSTM-RL	0.0323	0.143	0.0216	0.0583	0.1221	0.0364
TPGN	0.0331	0.149	0.0221	0.0612	0.1293	0.0379

The main evaluation results on the ICE dataset are reported in Table 2. The widely-used BLEU-4 [37], METEOR [38], CIDEr [27], ROUGE-L [39], and SPICE [40] metrics are reported in our quantitative evaluation of the performance of the baseline models. In evaluation, the first baseline, LSTM-XE, is the widely used CNN-LSTM method originally proposed in [2] where XE stands for cross entropy loss function. The second baseline, LSTM-RL, is the same as LSTM-XE except that the loss function is a negative reward and a reinforcement learning (RL) procedure is used for training the LSTM. The third baseline is TPGN (Tensor Product Generation Network) [31]. As shown in Table 2, the TPGN outperforms LSTM-XE and LSTM-RL in all metrics across the board. For comparison, we trained an LSTM with the COCO data set under the cross entropy loss function; then tested it with images in ICE and compared the generated sentence to the gold comments; its performance is listed in the second row (beginning with ‘Image captioning’) in Table 2.

Fig. 9 and Fig. 10 show examples of comments generated by baseline models and humans. It can be observed that the image captioner produces factual descriptions of images while an image commenter produces comments that are emotional, subjective, imaginative, and more likely to engage human interaction. For example, for the third image in Fig. 9, the image captioner generates “up a skateboard in the air” while the LSTM-XE produces “you gona sink”, since a human imagines that a low-altitude island is easy to sink under the sea level.

5.2 Interpretation

From the experiments described in Section 5.1, we observe that an image commenting system trained with our dataset achieves human-like understanding of events, and expresses itself with sentiments/subjectively to some extent. For the example of Fig. 7(a), the LSTM-XE generates “such a beautiful animal, wonderful world”, which is subjective and expresses human affects; in contrast, the image captioner produces “hand up to hit a tennis ball”, which describes an



Fig. 7. Example of Test cases

action or a fact. For the example of Fig. 7(b), the LSTM-XE generates “gorgeous shot”, which conveys the excitement like a human; in contrast, the image captioner produces “a young boy on a skateboard”, which is a dull fact.

Moreover, the ICE dataset holds a great potential of facilitating human machine interaction. For example, the commercial chat-bot including Microsoft chat-bot and Facebook messenger may benefit from the ICE dataset. Fig. 8 shows three examples of a commercial chat-bot interacting with humans.



Fig. 8. Three sessions of commercial chat-bot interacting with humans

6 Conclusions

In this paper, we define a novel image-to-text generation task, called image commenting (ICE), and provide the first dataset for this task. Our task moves from

image factual descriptions to comments that capture human imagination, emotion, and social interactions. We would like to emphasize that the social language captured in this dataset is critical for training conversational systems/machines, capable of human-like thinking, reasoning, and understanding. We have established several strong baselines for the task of image commenting. The dataset ICE v.1 has values in improving the commercial conversational system. We hope that this dataset also benefit more applications in visual dialog systems, visual question-answering systems, social chat-bots, intelligent personal assistants, and natural language processing systems.

				
	that man doe not age.	how doe it fly.	you gona sink.	that is so hideous.
LSTM-XE				
	the picture in the white.	i seeing the flying.	that is just low.	that is so good.
LSTM-RL				
	til clinton is immortal.	the coolest part of over there.	remind me of city.	like some food straight.
TPGN				
Human generated comments	1) 1955? 2) Sorry for the incorrect year in the title. It should say 1995. The picture was taken in the White House. 3) That man does not age. 4) TIL Bill Clinton is immortal. 5) I'm sorry to disturb your believe that this was taken in Oslo 1955. Because I can assure you it was not.	1) At a quick glance, there is an optical illusion that makes the vehicle look like a giant robot, that dwarfs the other 2 planes, and is crawling around the desert. 2) Neil Armstrong was nearly killed in one of them. Had to eject twice. 3) How does it fly? 4) Also known as the Flying Bedstead. 5) THIS GIANT MACHINE WILL KILL US ALL!	1) SimCity gone crazy - I love it! Would love to visit. 2) Man, that is gonna be one hel-luva scuba diving spot on 20 years 3) How do they anticipate for the sea levels rising ? 4) How do they do for sewage and water ?? O.O 5) Suddenly living in New Orleans doesn't seem so bad.	1) I'm sorry, but that is actually the ugliest building I've ever seen. 2) Oh brutalism, why do you hate pedestrians so much? 3) That looks like some shit straight outta 1984. 4) That is so hideous.

Fig. 9. Comments generated by baseline models and humans.

				
	that would be cool.	good food.	i am loving the new.	I like this so much.
	he did not know about the.	bacon look nice though there.	thought that we for a second then.	that is awesome.
	it a big boat and water.	picture with a potato.	more aggressive look but is putting out.	it is pretty interesting.
Human generated comments	1) TIL there were submarines way back in the 1860s... whoa. EDIT: Even though it didn't work, I didn't know about the Hunley. 2) that would be cool to find a shot of charleston today, in the same spot 3) So this attacked other ships by ramming a petard like charge-on-a-stick against the enemy? That must have been kind of scary.	1) I wish the picture had less, well, probably contrast. I can recognize your description but the picture looks... weird, sorry. Bacon looks crispy though. 2) Was this picture taken with a potato? 3) Shitty filter makes good food shitty.	1) I'm loving the new more aggressive look BMW is putting out. 2) thought that was forza for a second then.	1) You should check out World's Toughest Fixes on Netflix they have an episode where they replaced one of those. It was pretty interesting. 2) Wow that's awesome. Kind of hard to get a sense of scale though. Outage time at your plant? 3) love the steam 4) I have some photos of a turbine lift, I should share soon.

Fig. 10. Comments generated by baseline models and humans.

References

1. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). In: Proceedings of International Conference on Learning Representations. (2015)
2. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3156–3164
3. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works. arXiv preprint arXiv:1505.01809 (2015)
4. Chen, X., Zitnick, L.: Mind’s eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2422–2431
5. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 2625–2634
6. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3128–3137
7. Kiros, R., Salakhutdinov, R., Zemel, R.: Multimodal neural language models. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). (2014) 595–603
8. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
9. Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. CoRR **abs/1504.00325** (2015)
10. COCO: Coco dataset for image captioning. <http://mscoco.org/dataset/#download> (2017)
11. Searle, J.: Minds, brains and programs. Behavioral and Brain Sciences (1980) 417
12. Park, C.C., Kim, G.: Expressing an image stream with a sequence of natural sentences. In: Advances in Neural Information Processing Systems. (2015) 73–81
13. Huang, T.H.K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., et al.: Visual storytelling. In: NAACL. (2016)
14. Liu, Y., Fu, J., Mei, T., Chen, C.W.: Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks. In: AAAI. (2017) 1445–1452
15. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: EMNLP. (2014)
16. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. (2014) 3104–3112
17. Mathews, A.P., Xie, L., He, X.: Senticap: Generating image descriptions with sentiments. CoRR **abs/1510.01431** (2015)
18. Gan, C., Gan, Z., He, X., Gao, J., Deng, L.: Stylenet: Generating attractive visual captions with styles. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 955–964

19. Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L.: Generating natural questions about an image. *arXiv preprint arXiv:1603.06059* (2016)
20. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: *Proceedings of International Conference on Computer Vision (ICCV)*. (2015)
21. Ren, M., Kiros, R., Zemel, R.: Question answering about images using visual semantic embeddings. In: *ICML Deep Learning Workshop*. (2015)
22. NLTK: Natural language toolkit (nltk). <http://www.nltk.org> (2017)
23. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (2004)
24. Wu, Q., Shen, C., Liu, L., Dick, A., van den Hengel, A.: What value do explicit high level concepts have in vision to language problems? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016) 203–212
25. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016) 4651–4659
26. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L.: Semantic compositional networks for visual captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017)
27. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2015) 4566–4575
28. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2017)
29. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* **8**(3–4) (1992) 229–256
30. Pasunuru, R., Bansal, M.: Reinforced video captioning with entailment rewards. In: *Proceedings of Conference on Empirical Methods on Natural Language Processing*. (2017)
31. Huang, Q., Smolensky, P., He, X., Deng, L., Wu, D.: Tensor product generation networks for deep nlp modeling. *arXiv preprint arXiv:1709.09118* (2017)
32. Smolensky, P.: Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence* **46**(1-2) (1990) 159–216
33. Smolensky, P., Legendre, G.: The harmonic mind: From neural computation to optimality-theoretic grammar. Volume 1: Cognitive architecture. MIT Press (2006)
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016) 770–778
35. Mikolov, T., Chen, K., Corrado, G., Dean, J.: word2vec: Tool for computing continuous distributed representations of words. <https://code.google.com/archive/p/word2vec/> (2018)
36. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng,

- X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015) Software available from tensorflow.org.
37. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics (2002) 311–318
 38. Banerjee, S., Lavie, A.: 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, Association for Computational Linguistics (2005) 65–72
 39. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Proceedings of the ACL workshop, Association for Computational Linguistics (2004)
 40. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision, Springer (2016) 382–398