NICE: Neural Image Commenting Evaluation with an Emphasis on Emotion and Empathy

Kezhen Chen [§*], **Qiuyuan Huang** [‡*], **Daniel McDuff** [‡*], **Jianfeng Wang** [‡], **Hamid Palangi** [‡], **Xiang Gao** [‡], **Kevin Li** [†], **Kenneth Forbus** [§], **Jianfeng Gao** [‡]
[‡]Microsoft Research, Redmond, WA;
[§]Northwestern University, Evanston, IL; [†]University of Michigan, Ann Arbor, MI
[‡]{qihua,damcduff,jianfw,hpalangi,xiag,jfgao}@microsoft.com,
[§]kzchen@u.northwestern.edu, [†]kevyli@umich.edu, [§]forbus@northwestern.edu

Abstract

Emotion and empathy are examples of human qualities lacking in many human-machine interactions. The goal of our work is to generate engaging dialogue grounded in a user-shared image with increased emotion and empathy while minimizing socially inappropriate or offensive outputs. We release the ***N**eural **I**mage **C**ommenting **E**valuation (**NICE**)* dataset consisting of almost two million images and their corresponding, human-generated comments, as well as a set of baseline models and over 28,000 human annotated samples. Instead of relying on manually labeled emotions, we also use automatically generated linguistic representations as a source of weakly supervised labels. Based on the annotations, we define two different task settings on the NICE dataset. Then, we propose a novel model - ***M**odeling **A**ffect **G**eneration for **I**mage **C**omments (**MAGIC**)* - which aims to generate comments for images, conditioned on linguistic representations that capture style and affect, and to help generate more empathetic, emotional, engaging and socially appropriate comments. Using this model we achieve state-of-the-art performance on one setting and set a benchmark for the NICE dataset. Experiments show that our proposed method can generate more human-like and engaging image comments.

1 Introduction

Recent progress in the field of natural language processing (NLP) and computer vision (CV) has led to considerable advances in the domains of image captioning, visual question answering, visual dialog and visual storytelling (Mao et al., 2015; Vinyals et al., 2015; Devlin et al., 2015; Chen and Zitnick, 2015; Donahue et al., 2015; Karpathy and Fei-Fei, 2015; Huang et al., 2018; Kiros et al., 2014a,b; Gao et al., 2019; Shum et al., 2018). Most image captioning systems focus on generating literal descriptions of content either directly or in the form of Q&A. Despite remarkable progress, developing intelligent dialogue agents that are capable of engaging in socially appropriate and empathetic conversations with humans is still very challenging. Fig. 1 shows examples of two images with comment threads. The caption for the first image generated by a captioning model is "Some houses are at the foot of a mountain". While this faithfully describes the image, imagine you posted the picture on social media and someone responded with that statement. Would that spark an engaging conversation or feel like an empathetic response? Probably not. A conversation is grounded not only in visible objects (e.g., houses and mountains) but also in events, actions and emotions (e.g., amazement at the grandeur of the mountain or a desire to climb it). It is the latter that are often as important in meaningful conversations and especially in forming emotional connections.

[*]Equal Contribution.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

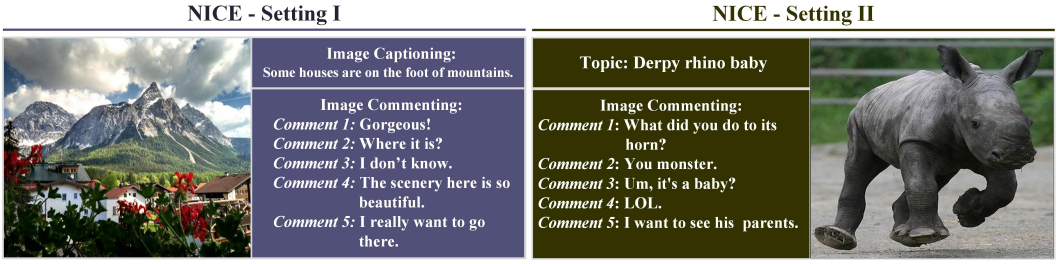


Figure 1: We present a dataset-NICE, and a novel approach MAGIC for generating comments to user shared images. In NICE-Setting I: In contrast to traditional image-captioning and image-grounded dialogue tasks we focus on synthesizing content that is empathetic, emotional and engaging. NICE-Setting II: Samples of NICE-Setting II Dataset with Topic.

In this work, we design a dialogue system that is capable of commenting on images in an emotional and engaging manner. To create a holistic measure of the performance of the models we selected five dimensions that capture different conversational qualities: empathy, emotion, engagement, social appropriateness and relevance to the topic. We make the assumption that it is desirable for automatically generated dialogue to score well across all of these measures. It is helpful to define the important terms in our work. Emotion here is defined as the use of language that refers to, or reflects, affect and is a response to a specific stimulus (in this case the image and/or other comments). This is differentiated from mood which is affect not related to a specific stimulus but capturing a longer lasting feeling that might influence a whole conversation. Empathy is defined as the ability to understand and share the feelings of another.

To summarize, the core contributions of this paper are: 1) Collecting and releasing a large dataset[2], NICE, which contains almost two million images and more than six million groups of comment dialogue conversation. 2) Defining two different task settings on the NICE dataset including a sizable manually and automatically annotated portion. 3) Providing a benchmark results using established metrics (e.g., BLEU, CIDEr) and via human judgements of empathy, emotionality, engagement, social appropriateness and relevance. 4) We also introduce a novel approach, MAGIC, to simulate human commenting on NICE dataset, which aims to generate targeted comments on a given image weakly supervised by affect features. Experiments show that MAGIC outperforms baseline methods on the NICE task.

2 Related Work

With the recent advances in deep learning, a growing number of researches are interested in studying vision and language jointly. Vision-language understanding has become one of the key components of conversational agents, such as Xiaoice (Weitz, 2014). A great deal of focus has been paid to image captioning (Lin et al., 2014; Sharma et al., 2018; Young et al., 2014), which typically focuses on literal descriptions of image content. However, in social conversations, people usually engage with others using language with emotions, opinions and subjectivity. For example, image commenting on social media has rich stylistic features. In this paper, we introduce the image comment generation task, where the aim is to build models that produce more engaging comments grounded in visual images. Specifically, we present a pre-training model for this task.

There are several pre-trained models that address various tasks across the language and vision space. Large-scale pre-trained models have achieved state-of-art results on many natural language processing and generation tasks (Peters et al., 2018; Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019; Radford et al., 2019). Pre-trained models learn representations using tasks such as predicting words based on their context. GPT-2 and CTRL are examples of language generation models that leverage pre-training.. We use a well validated linguistic style representation to control our Magic model. We extract affect features for auto-labeling which used to learn a control input related to word categories. Some researches have also combined vision and language features in pre-trained models for various downstream vision-language tasks (Lu et al., 2019; Tan and Bansal, 2019; Zhou et al., 2019; Chen

[2]Code and Data: https://github.com/ckzbullbullet/NICE.

et al., 2019; Alberti et al., 2019; Li et al., 2019a,b). One of the closest pre-trained generation models that compare with our work is the unified vision language pre-training (VLP) model. However, VLP focuses on generating image captions and lacks the ability to generate expressive, stylistic responses. To alleviate this problem, we propose our MAGIC pre-traing model to fill this gap and the proposed Image Commenting task offers a more natural setting for generating and evaluating affect dialogue about visual content.

3 NICE Dataset

3.1 Dataset Construction

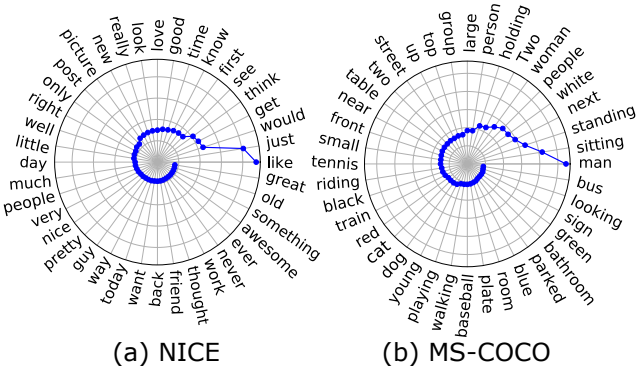
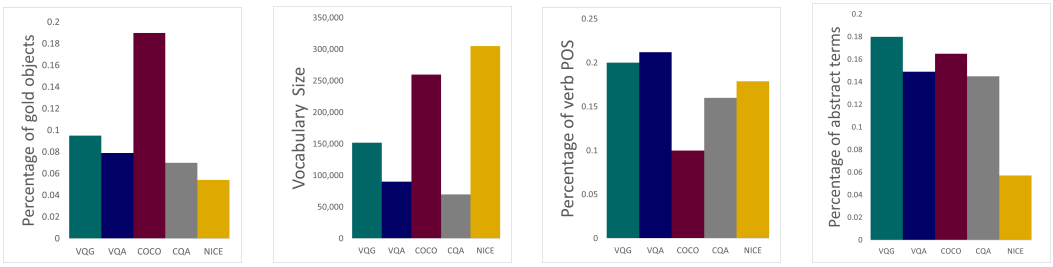


Figure 2: Frequency of the top 40 words in the (a) NICE and (b) COCO datasets. The radius reflects the frequency of the corresponding word (larger radius = higher the frequency).

The NICE dataset consists of over 2M images, and 7M image-comment pairs split into training, validation, and test sets. In this section, we first describe how the dataset was collected, and then present some of its unique characteristics. Our goal is to simulate natural comments from humans, which is a task that requires a large volume of data. Therefore, we scraped 10 million image-comment pairs from Reddit[3]. Each thread was required to start with an image and contain at least one comment. We applied filters to both the images and comments to remove sensitive content such as people's names, adult or pornographic content, politically sensitive language, ethnic-religious content, or other potentially offensive or contentious material (including inappropriate references to violence, crime and illegal substances). After filtering, the number of images of the dataset was reduced to 2,233,926 samples and the number of image-comment pairs was reduced to 7,304,680 samples. This filtering was performed with pre-trained classifiers. We believe that this dataset presents a valuable resource for the community. Below we highlight some of the properties of the data.

3.2 Dataset Properties



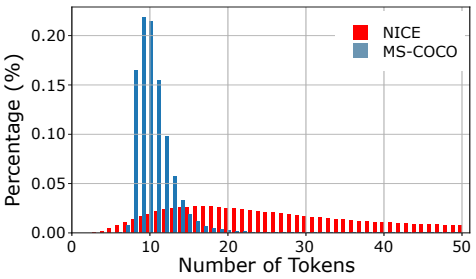
(a) (b) (c) (d)

Figure 3: Comparison of annotations on the NICE dataset: (a) % of human-labeled objects used in annotations, (b) vocabulary size, (c) % of verb POS, (d) % of abstract terms.

High-Frequency Words. First, we list the 40 highest-frequency words in the NICE dataset and compare these to the top words in the captions from the COCO dataset (Lin et al., 2014). As shown in Fig. 2, there is almost no overlap among the lists from the two datasets. Showing that the types

[3]https://www.reddit.com/

of language used in image commenting is quite different from that used in image captioning - this reinforces our decision to collect these data.



(Fig. 4)

Figure 4: Histogram of the length of sentences in NICE dataset and COCO dataset.

Comparison of Various Annotations. Fig. 3 shows summary statistics for several image-to-text datasets. Fig. 3 (a) compares the percentage of gold object-mentions in each of the annotations. Object-mentions are the words associated with the human-labeled object boundary boxes as provided in the COCO dataset. As reported in VQG (Mostafazadeh et al., 2016), COCO captions have the highest percentage of these literal objects. Because object-mentions are often the answers to the questions in VQA (Antol et al., 2015) and CQA (Ren et al., 2015), those questions naturally contain objects less frequently. On the contrary, comments in the NICE dataset have the lowest percentage of human-labeled objects, as comments are less descriptive and more about expressing opinions, sentiment, and emotion. Fig. 3 (b) shows that the NICE dataset has the largest vocabulary size. This is expected due to the large number of comments (7M) and the fact that comments in social chats tend to be more diverse. Fig. 3 (c) shows that verbs represent a high percentage of words in the NICE dataset. Fig. 3 (d) indicates that the NICE dataset uses significantly less abstract terms than the other datasets. These analyses show that the NICE dataset, though also focused on image-to-text generation, has very different properties from the other datasets.

Length of sentences. Fig. 4 shows a histogram of the number of tokens in the text from the NICE and COCO datasets. On average comments in NICE are longer (38.43 tokens) than captions in COCO (10.46 tokens); but more significantly, the comments have much larger variance in length. The COCO captions were created under conditions with clear guidelines about the nature of the descriptions. The NICE data contains examples more akin to free-form comments.

Sentiment Words. We find that 11 of the top 40 most frequent words had non-neutral sentiment, as shown in Table 1 of the Appendix. The sentiment labels were generated using an off-the-shelf sentiment analysis tool NLTK (Toolkit, 2017). Readers are referred to the Appendix 1 for more comparative analysis of sentiment words in the NICE dataset.

4 NICE-Setting I (Human Labeling)

4.1 Human Labeling for NICE-Setting I

For some qualities (e.g., empathy or social appropriateness) there are currently no automated metrics for evaluating dialogue generation models. These qualities are of particular importance in our task. Therefore, we had human labelers code a large set (over 28,000) of images and comments. These samples form the validation and testing sets of our dataset. During each Human Intelligence Task (HIT) we showed a labeler an image accompanied by a comment from a single thread associated with the image. As a single image can have multiple comment threads we randomly selected one comment thread for each image per HIT. The labeler was asked to rate how socially appropriate, empathetic, emotional and relevant to the image the comments were. Each rating was performed on a scale of 1 (not at all) to 7 (extremely). They were also asked whether the text featured offensive content (No/Yes). In total, 28,392 image and comment samples were labeled. Each sample was labeled by one labeler, but due to the large number of samples we had a total of 180 labelers, each who labeled an average of 156 images. The complete set of labels are included in the dataset.

4.2 Experiments on NICE-Setting I

We split the NICE dataset, described in Sec. 3, into training (1,908,902 image-comment pairs), validation (human labeling; 13,896), and testing (human labeling; 14,496) sets. The data split

	Automatic Metrics				Human Manual Evaluation				
Methods (%)	**Bleu-4**	**Rouge**	**Cider**	**Spice**	**Engag.**	**Emo.**	**Empath.**	**Appro.**	**Relev.**
LSTM-XE	0.29	8.60	1.74	1.40	3.39 (.21)	3.07 (.27)	3.29 (.23)	3.78 (.25)	3.81 (.26)
Caption-Bot	0.30	8.20	3.20	2.00	3.53 (.22)	3.14 (.29)	3.13 (.22)	3.97 (.26)	4.52 (.23)
SCN	0.30	8.40	1.70	1.50	3.53 (.23)	2.99 (.28)	3.01 (.23)	3.95 (.27)	3.94 (.27)
BUTD	0.78	10.31	1.52	1.00	3.44 (.21)	3.33 (.28)	3.40 (.24)	3.93 (.27)	3.95 (.27)
VLP	0.80	10.40	3.20	1.50	3.79 (.19)	3.45 (.28)	3.51 (.22)	4.22 (.23)	4.52 (.23)
Human	-	-	-	-	4.53 (.20)	4.09 (.23)	4.41 (.20)	4.85 (.21)	5.13 (.21)

Table 1: Performance on the NICE-Setting I dataset. Left) Automatic metrics. Right) Human evaluation. Performance on the ground-truth (human) comments shows a empirical limit on the scores. Numbers in brackets reflect standard errors. We showed previous state-of-the-art methods: LSTM-XE (Vinyals et al., 2015), Caption-Bot (Microsoft, 2017), SCN (Gan et al., 2017), BUTD (Anderson et al., 2018), VLP (Zhou et al., 2019).

will be released along with the dataset. For LSTM based baselines (LSTM-XE, SCN, BUTD), we used a vocabulary that consists of 18,018 words. For Transformer based model (VLP) we used a vocabulary of size 28,996. In all the experiments, for CNN based baselines (LSTM-XE, SCN) we used ResNet-152 (He et al., 2016), pretrained on the ImageNet dataset, to extract image features. For object detection based baselines (BUTD and VLP) we used an object detector pretrained on the visual genome dataset with 1,600 object classes. The feature vector $\boldsymbol{v}$ is of size 2048.

Baseline Models on NICE-Setting I. Now let us compare the baseline models we used to evaluate performance on the proposed NICE task. This is important to provide a comprehensive picture of the current performance of state-of-the-art methods on the NICE task. The details of the baseline models can be found in the Appendix 2.1.

Automatic Evaluation. The BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), ROUGE-L (Lin, 2004), and SPICE (Anderson et al., 2016) evaluation results are reported in Table 1. The results shows that the baseline models, including state-of-the-art image captioning models such as BUTD (Anderson et al., 2018), perform relatively poorly.

Human Evaluation. We had 200 images and the corresponding generated comments from each model annotated by human labelers. We used the same procedure as the annotation described in Sec. 4.1. The humans rated each generated comment in terms of how engaging, emotional, empathetic, appropriate and relevant it was. Table 1 shows the average scores for each model on these metrics. The VLP model produced comments that were rated as more engaging (μ=3.79), emotional (μ=3.45), empathetic (μ=3.51) and appropriate (μ=4.22) than other baselines. The model can't captures the overarching emotional tone of the dialog more effectively as human. The responses were rated as less relevant than captions generated using an image captioning model. This is expected as the image captioning model output references specific objects in the image, where as emotional content is by nature more abstract. It is challenging for the dialogue to satisfy all criteria but we believe there is scope for improvement over our baseline.

5 NICE-Setting II (Auto-Labeling)

Based on setting I, we also have another setting II.The input in this case is an image, the thread title and the current comment history. We applied similar filters as in setting I on the image and text. In setting II, we treat the title of the thread as the "comment topic". When people provide the comments on an image, they express their perspectives based on their affective state, the comment topic and the information from the comment history. After the filtering, the dataset finally has 2,150,528 images and 6,720,542 comment dialogue threads, where each dialogue has a thread topic and up to five comments like the sample in Fig. 1. In this section, we first introduce the unique characteristics of the Setting II and introduce a new model, Modeling Affect Generation for Image Commenting (MAGIC), for image commenting on Setting II.

5.1 Affect Features

For each comment in a thread, affect features are extracted to represent the language style and emotions. To replace manual annotation, and capture the rich information in the comments, we select Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) to represent the affect and style features of the comments. LIWC is widely used for text analysis in linguistic field and psychology, and has been demonstrated to capture important information (Chung and Pennebaker, 2018). In this paper, we utilized the LIWC 2007 dictionary, which was composed of 2,290 words and word stems, and each word or word stem defines one or more word categories or sub-dictionaries. With the LIWC tool, we extract a 64-dimension feature vector for each comment automatically and this vector is normalized. We hypothesize that these features can represent the open-domain human affect and language style in the comments.

5.2 Definition of Affect Image Commenting Task on NICE -Setting II

We define the MAGIC task as generating comments in response to a shared image, similar to a dialog response in a social conversation setting in order to maximize user engagement and eventually form long-term, emotional connections with users. We formalize the generation task as follows: each sample of this dataset has an image $\boldsymbol{I}_{image}$, a comment topic $\boldsymbol{H}$ of the whole dialogue, and N comments $\boldsymbol{C}_1, ..., \boldsymbol{C}_N$ with corresponding thread affect distribution features $\boldsymbol{A}_1, ..., \boldsymbol{A}_N$. Systems aim to construct a plan to generate the comment $\boldsymbol{C}_i$ using the current state information $S_{I,T,i-1}$, which contains the input image features $\boldsymbol{I}$, comment topic $\boldsymbol{H}$, and the comments history $(\boldsymbol{C}_1, ..., \boldsymbol{C}_{i-1})$, and is conditional on the affect feature $\boldsymbol{A}_i$, which represented as a affect feature vector $\boldsymbol{P}_i$.

6 MAGIC Model on NICE-Setting II

Following the success of large-scale pre-training, we introduce a novel model, Modeling Affect Generation for Image Commenting (MAGIC), which aims to generate emotional comments conditioned on an image, a comment topic, affect features, and the comment history. We will introduce MAGIC model and our training procedure in the following.

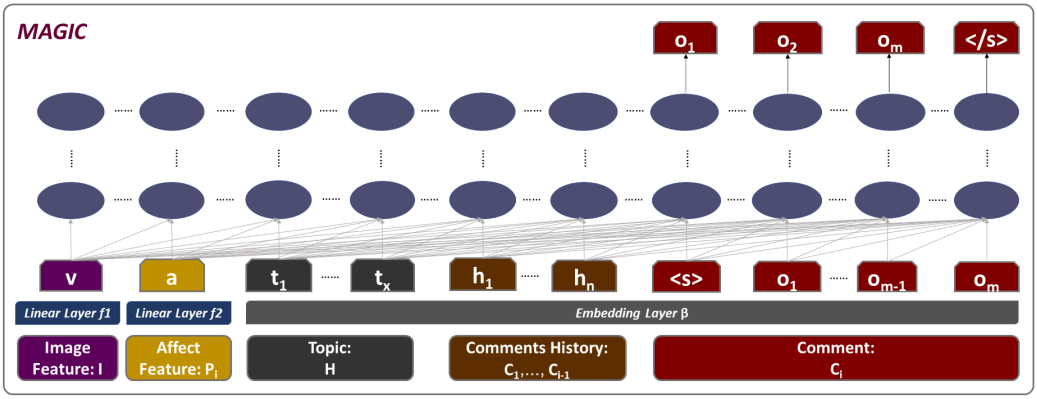


Figure 5: Overview of MAGIC model architecture

6.1 MAGIC Training on NICE-Setting II

As large models usually generalize better to new domains when they are trained on large volumes of data, we extend GPT-2 (Radford et al., 2019) as the backbone to the language generation in our MAGIC model. GPT-2 is a transformer-based language model trained on large scale web data and uses self-attention where each token attends to its left tokens. It is trained with the objective: predict the next word, given all of the previous words within a defined context window. We trained MAGIC with the same stage as the small-sized GPT-2 model, which has 12 layers and each layer has 12 heads. Based on the definition in 5.2, the model aims to compute the conditional probability L:

$$L = p(\boldsymbol{C}_i | \boldsymbol{I}, \boldsymbol{H}, \boldsymbol{P}_i, \boldsymbol{C}_1, ..., \boldsymbol{C}_{i-1}) \quad (1)$$

In MAGIC training, as showed in Fig 5, we encode the input image into a 2048-dimension feature vector $\boldsymbol{I}$ using pre-trained Resnet-152 model (He et al., 2016). The affect and style feature $\boldsymbol{A}_i$ (introduced in 5.1) is represented as a affect feature vector $\boldsymbol{P}_i$. The image feature vector $\boldsymbol{I}$ and affect feature vector $\boldsymbol{P}_i$ are passed to two separate linear layers f_1, f_2 to map to two 768-dimension vectors $\boldsymbol{v}$ and $\boldsymbol{a}$. Then, comment topic $\boldsymbol{H}$, history comments $(\boldsymbol{C}_1, ..., \boldsymbol{C}_{i-1})$ and output comment $\boldsymbol{C}_i$ are fed into an embedding layer β to generate embedding vectors for each token respectively, $\boldsymbol{t}_1, ..., \boldsymbol{t}_x$, $\boldsymbol{h}_1, ..., \boldsymbol{h}_n$ and $\boldsymbol{o}_1, ..., \boldsymbol{o}_m$ for each token as following:

$$\boldsymbol{v} = f_1(\boldsymbol{I}), \qquad \boldsymbol{a} = f_2(\boldsymbol{P}_i) \tag{2}$$

$$E_{topic} = \boldsymbol{t}_1, ..., \boldsymbol{t}_x = \beta(\boldsymbol{H}) \tag{3}$$

$$E_{history} = \boldsymbol{h}_1, ..., \boldsymbol{h}_n = \beta(\boldsymbol{C}_1, ..., \boldsymbol{C}_{i-1}) \tag{4}$$

$$E_{comment} = \boldsymbol{o}_1, ..., \boldsymbol{o}_m = \beta(\boldsymbol{C}_i) \tag{5}$$

The encoded image feature vector $\boldsymbol{v}$, the affect feature vector $\boldsymbol{a}$, the embedded comment topic vector $\boldsymbol{t}_1, ..., \boldsymbol{t}_x$, the embedded history comments vectors $\boldsymbol{h}_1, ..., \boldsymbol{h}_n$ and the embedded output comment vectors $\boldsymbol{o}_1, ..., \boldsymbol{o}_m$ are concatenated together as following:

$$\boldsymbol{B} = f_{concat}(\boldsymbol{v}, \boldsymbol{a}, E_{topic}, E_{history}, E_{comment}) \tag{6}$$

Then, $\boldsymbol{B}$ is fed to MAGIC model for training. For each transformer head, we use the masked version of the self-attention on query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$ and value matrix $\boldsymbol{V}$ with mask matrix $\boldsymbol{M}$ as following:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{M} \circ \boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V} \tag{7}$$

The prediction loss is only computed for $\boldsymbol{o}_1, ..., \boldsymbol{o}_m$.

6.2 Inference and Learning Strategy

Firstly, we formalize the training procedure. Given a training dataset with D samples, all comments in each sample have total Y tokens. We use maximizing the log-likelihood (MLE) to learn the model parameters θ of the conditional probabilities L_θ over the entire training dataset:

$$B^{i,m} = f_{concat}(\boldsymbol{v}^i, \boldsymbol{a}^i, E^i_{topic}, E^i_{history}, \boldsymbol{o}^i_1, ..., \boldsymbol{o}^i_m) \tag{8}$$

$$L_\theta(D) = \sum_{i=1}^{D} \sum_{m=1}^{Y} p_\theta(\boldsymbol{o}^i_m | B^{i,m-1}) \tag{9}$$

During inference, each token is generated one by one via beam search with beam size 2.

7 Experiments of MAGIC

We split the subset of NICE data with 6,550,542 image-comment pairs for training, 100,000 image-comment pairs for validation, and 70,000 image-comment pairs for testing. We trained MAGIC 30 epochs with batch size 36 on each GPT using a machine with 4xV100 32G GPUs and the learning rate was $5e-5$. For the baseline models, we modified two off-the-shelf image-captioning models, Show Attention and Tell (ShowAttTell) (Xu et al., 2015) and Bottom-Up-Top-Down Attention (BUTD) (Anderson et al., 2018), for the same task setting as MAGIC and compared with our model on NICE dataset. Details about modifying baseline models are described in Appendix.

Table 2 shows the performance of our MAGIC model and previous state-of-the-art methods on the NICE dataset. To evaluate the performance of the MAGIC model and whether affect features provide rich information for comment generation, we evaluate three different aspects of the generated comments: token matching, embedding similarity and diversity. For token matching metrics, MAGIC outperforms ShowAttTell and BUTD on all four metrics. As users' comments can have different words with similar affect, we also utilize the SPICE (Anderson et al., 2016) and Bert-Score (Zhang

et al., 2019), which have been widely used for embedding similarity. Results show that MAGIC has higher performance on both scores (Zhang et al. (2019) recommends to use BertF1 for comparison). Finally, we tested the diversity of generated comments. We tested Entropy4 and Distinct2 from Qin et al. (2019). As MAGIC is pre-trained on large volume of data, they have higher diversity than ShowAttTell and BUTD. Figure 6 shows some generated comment samples from MAGIC model comparing with generated samples from BUTD model.

	Token Matching				Embedding Similarity				Diversity	
Model	**Bleu1**	**Bleu4**	**ROUGE**	**CIDEr**	**SPICE**	**BertP**	**BertR**	**BertF1**	**Entropy4**	**Distinct2**
ShowAttTell-Affect	0.274	0.050	0.227	0.579	0.053	0.227	0.146	0.184	10.201	0.126
BUTD-Affect	0.299	0.056	0.269	0.763	0.064	**0.249**	0.134	0.189	9.851	0.043
GPT-2-NoAffect	0.065	0.003	0.056	0.051	0.011	0.040	0.037	0.037	12.706	0.211
MAGIC (ours)	**0.306**	**0.062**	**0.288**	**0.852**	**0.071**	0.204	**0.203**	**0.202**	**13.709**	**0.297**

Table 2: Automatic Evaluation results of four models on NICE dataset. Comparing with ShowAttTell (Xu et al., 2015) and BUTD (Anderson et al., 2018), MAGIC outperforms the other models in token matching, embedding similarity and diversity.

Figure 6: Generated comment samples using MAGIC model on NICE-Setting II.

8 Conclusion and Future Work

In this paper, we present a new vision-language task called Neural-Image-Commenting-Evaluation (NICE) which extends image descriptions to comments with an emphasis on emotion and empathy. We design two task settings on this dataset based on different annotations. For NICE-setting II, we propose a novel large-scale model, MAGIC, for image commenting conditional on affect and style features. Comparing with other model, MAGIC has better ability to generate affective and emotional image comments. To facilitate research in this area, we will release the NICE dataset. The social language captured in this dataset is of great value for training conversational systems to imitate human-like thinking, reasoning, and understanding. Image commenting is an emerging area and there is much room for future work. While we anticipate that the task we are proposing can have a significant positive impact in many domains (e.g., accessibility, storytelling, entertainment), we acknowledge that they can be abused (e.g., fake comment generation) and countermeasures may need to be developed. We hope that solving the NICE task will benefit a wide range of applications including visual dialogue generation, visual question-answering and help create better social chat-bots and intelligent personal assistants.

Acknowledgement

We are especially grateful to Harry Shum, Lei Zhang, Xiaodong He for their comments, suggestions, and painstaking multiple reviews of this paper, and their pointers to the literature. We thank Zhe Gan for his work and his generous feedback for the project. We are grateful to Mary Czerwinski for her enormous support and encouragement. We thank all the XiaoIce teams for their supporting the work during the early stage.

References

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*.

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 International Conference on Computer Vision (ICCV)*.

Xinlei Chen and Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2422–2431.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Universal image-text representation learning. *arXiv preprint arXiv:1909.11740*.

Cindy K. Chung and James W. Pennebaker. 2018. What do we know when we liwc a person? text analysis as an assessment tool for traits, personal concerns and life stories. *The SAGE Handbook of Personality and Individual Differences*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and Trends in Information Retrieval*, 13(2-3):127–298.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Qiuyuan Huang, Pengchuan Zhang, Oliver Wu, and Lei Zhang. 2018. Turbo learning for captionbot and drawingbot. In *Proceedings of Neural Information Processing Systems (NeurIPS)*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014a. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014b. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*.

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

Chin-Yew Lin. 2004. Rouge: A package for tomatic evaluation of summaries. In *Proceedings of the ACL workshop*. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proceedings of International Conference on Learning Representations*.

Microsoft. 2017. Captionbot and captionbot api. `https://www.captionbot.ai/`.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading. *arXiv preprint arXiv:1906.02738*.

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

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Question answering about images using visual semantic embeddings. In *ICML Deep Learning Workshop*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

The Natural Language Toolkit. 2017. Sentiment analysis tool. *http://www.nltk.org/howto/sentiment.html*.

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.

Stefan Weitz. 2014. Meet xiaoice, cortana's little sister. *Bing Blogs*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*.