NICE: Neural Image Commenting with Empathy

Kezhen Chen[‡§], Qiuyuan Huang[‡], Daniel McDuff[‡],
Xiang Gao[‡], Hamid Palangi[‡], Jianfeng Wang[‡], Kenneth Forbus[§], Jianfeng Gao[‡]
[‡]Microsoft Research, Redmond; [§]Northwestern University, Evanston
kci1962@ads.northwestern.edu,
{qihua,damcduff,xiag,jianfw,jfgao}@microsoft.com
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 *Neural **I**mage **C**ommenting with **E**mpathy (**NICE**)* dataset consisting of almost two million images and the corresponding human-generated comments, a set of human annotations, and baseline performance on a range of models. Instead of relying on manually labeled emotions, we also use automatically generated linguistic representations as a source of weakly supervised labels. Based on these annotations, we define two different tasks for the NICE dataset. Then, we provide a novel pre-training 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 of our NICE tasks. The experiments show that the approach 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; Kiros et al., 2014a,b; Gao et al., 2019; Shum et al., 2018). Most image captioning tasks focus on generating literal descriptions of content either directly or in the form of searching or understanding.

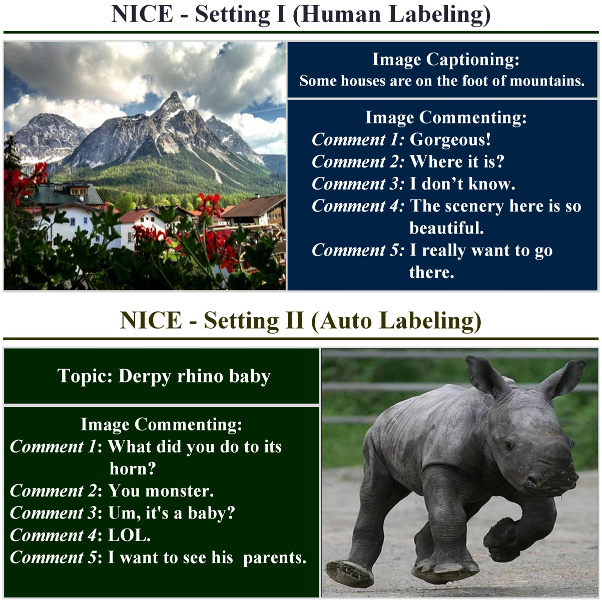

NICE - Setting I (Human Labeling)
Image Captioning:
Some houses are on the foot of mountains.
Image Commenting:
Comment 1: Gorgeous!
Comment 2: Where it is?
Comment 3: I don't know.
Comment 4: The scenery here is so beautiful.
Comment 5: I really want to go there.
NICE - Setting II (Auto Labeling)
Topic: Derpy rhino baby
Image Commenting:
Comment 1: What did you do to its horn?
Comment 2: You monster.
Comment 3: Um, it's a baby?
Comment 4: LOL.
Comment 5: I want to see his parents.

Figure 1: We present a dataset-NICE and a novel pre-training model-MAGIC for generating comments for user shared images. There are two examples for two NICE dataset Settings. 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: The second setting aims to generate dialogue-style comments based on a comment topic and affect features.

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 the examples of two images with comment threads for two NICE-Settings. The caption for the first image generated by a captioning model of the NICE-Setting I is “Some houses are at the foot of a mountain”. While this somewhat 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). Emotions are important in meaningful conversations and especially in forming emotional connections. Generating emotional comments would imitate human-like behavior, which is essential to human-machine interaction and conversation.

In this work, we present the Neural Image Commenting with Empathy (NICE) dataset with two task settings, and 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 for image commenting systems, we selected five dimensions that capture different conversational qualities: empathy, emotion, engagement, social appropriateness and relevance to the image-commenting pairs. We make the assumption that it is desirable for automatically generated dialogue to score well across all of these measures. Emotion and empathy in comments are specified. 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. We believe that this task will benefit various research fields such as vision-language and human-machine interaction.

To summarize, the core contributions of this paper are: 1) We collect and release a large dataset[1], NICE, which contains almost two million images and more than seven million groups of comment dialogue conversation. 2) We define two different tasks on the two NICE dataset settings including a sizable manually and automatically annotated portion. 3) We provide 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 pre-training 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-Setting II.

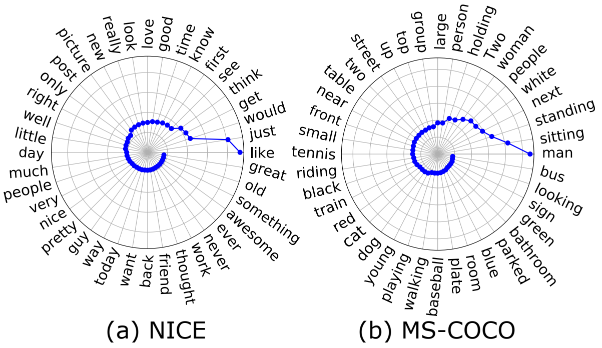
2 Related Work

With the recent advances in deep learning, a growing number of researchers are interested in studying vision and language jointly. Vision-language understanding has become one of the key components of conversational agents, such as Cortana (Microsoft, 2014). A great deal of focus has been paid to image captioning (Lin et al., 2014; Sharma et al., 2018a; 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 human-machine interaction system 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 the comment generation. We extract affect features for auto-labeling which are 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 et al., 2019; Alberti et al., 2019; Li et al., 2020, 2019). 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-training model to fill this gap and the proposed Image Commenting task offers a more natural setting for generating and evalu-

[1] Users can access code and to download the dataset at our official website: `https://nicedataset.github.io/`. Use of the code and dataset are governed by an End User License Agreement (EULA) to avoid any potential violation of rights or terms of service.



(a) NICE (b) MS-COCO

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).

ating comment dialogue with affection about visual content.

3 NICE Dataset

3.1 Dataset Construction

The NICE dataset consists of over 2M images, and 7M image-comment pairs (English) split into training, validation, and testing 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 requires a large volume of data. Therefore, we scraped 10 million image-comment pairs from website. Each thread was required to start with an image and at least one comment. We applied filters to both the images and comments to remove sensitive content such as adult or pornographic content, racy and gory content, non-English language, ethnic-religious content, and some sensitive content (including people's name, documents invoices, bills, financial reports) or other potentially offensive or contentious material (including inappropriate references to violence, crime and illegal substances). This filtering was performed with several open-domain API. For example, we used the "Microsoft Adult Filtering API" (Microsoft, 2019) to remove adult, racy and gory images, we use the "Detecting image types API" (Microsoft, 2018) to remove clip art and line drawings, we use the "Optical Character Recognition (OCR) API" (Microsoft, 2020) to remove printed or handwritten text from the images, such as photos of license plates or containers with serial numbers, as well as from documents invoices, bills, financial reports, articles, and more. We also removed people's names, politically sensitive language, ethnic-religious content, or other potentially offensive material (including inappropriate references to violence, crime and illegal substances) as the similar filter API for language cleaning. The last step of filtering, we make sure that NICE dataset had no more than 5 ($\leq$ 5) corresponding comments for each image, and there are not more than 6 ($\leq$ 6) different dialogue threads for the same image. In NICE-Setting II, after annotation, we filter out image-comments pairs without affect feature or dialogue topic from dialogue thread. We will keep cleaning and maintaining it in future.

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. Refer to Appendix A to find the details of dataset cleaning.

We believe that this dataset presents a valuable resource for the community. Below we highlight some of the properties of the data.

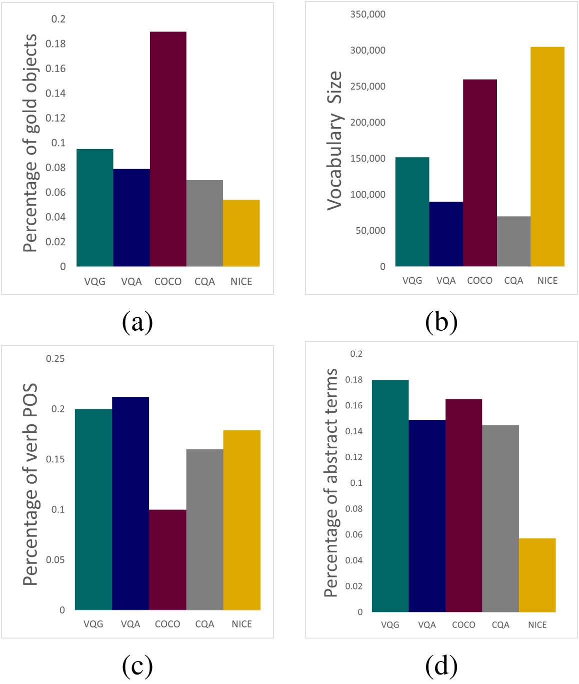
3.2 Dataset Properties

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. This observation reveals that the types of language used in image commenting are quite different from those used in image captioning, which reinforces our decision to construct the 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

Table 1: **Frequency of sentiment words in NICE**

Sentiment words	like	love	good	great	beautiful	pretty	nice	amazing	awesome	right	gorgeous
Frequency	**0.0071**	**0.0028**	**0.0025**	**0.0021**	**0.0019**	**0.0018**	**0.0017**	**0.0015**	**0.0013**	**0.0013**	**0.0009**



(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.

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 fewer abstract terms such as "think" or "win" than the other datasets. Following Mostafazadeh et al. (2016), we use a list of most common abstract terms in English (Vanderwende et al., 2015). The result is expected because sentences in the NICE dataset are more likely in the colloquial language style, which is often the case for engaging in social media. 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.

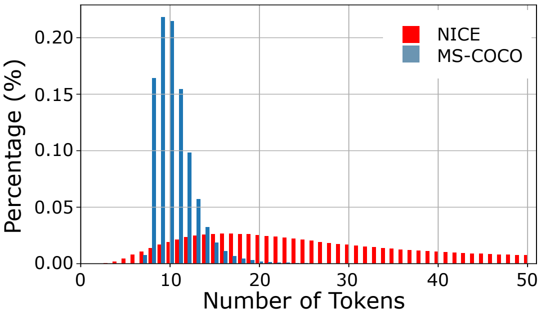


Figure 4: Histogram of the length of sentences in NICE dataset and COCO dataset.

Sentiment Words. Following Hu and Liu (2004), we extracted top 40 sentiment words for NICE dataset as shown in Table 1.

The most popular word in the NICE dataset is "like", which is a word with strong positive sentiment. Referring to the sentiment word list from Hu and Liu (2004), we find that 11 words among the top 40 words are sentiment words, as shown in Table 1 as below. Interestingly, all the 11 words express positive sentiment. This also reveals a bias in the real scenario: the usual comments tend to be of a positive sentiment or people are likely to show a positive attitude in conversations. On the contrary, the most frequent words in the COCO dataset tend to be the ones that describe facts such as action or objects, and do not contain any sentiment words listed in Hu and Liu (2004). The sentiment labels are generated using an off-the-shelf sentiment analysis tool NLTK (Toolkit, 2017). This demonstrates that the comments in the NICE dataset often contain opinions, emotional and subjective expressions, description of subjects, events, and scenes with unbounded scope, while the captions in the COCO dataset are more factual-oriented descriptions of images.

4 NICE Dataset Settings

4.1 NICE-Setting I (Human Labeling)

The NICE-Setting I of the dataset has over 28,000 human annotated samples. The top sample in Fig. 1 shows an example of this NICE-Setting I.

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. However, these qualities are particularly important for our data 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 NICE-Setting I. 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. A screenshot of the labeling task is shown in Appendix B. Each HIT involved viewing an image and six associated comments in the sequence that they were posted. The labeler was asked to rate how socially appropriate, empathetic, emotional, engaging 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). After that, we use the "Heuristic for filtering" algorithm (appendix.B) as a criteria to filter as constituting a “clean" human labeling dataset. The percentage of comments labeled as offensive was 3.2% (902/28392). While this might seem small, our labeling also captured whether comments were appropriate and comments that were not deemed offensive could be labeled as inappropriate. A further 8.1% were deemed inappropriate on a scale of 0 (inappropriate) - 1 (appropriate). 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. We compensated labelers at a calculated rate of $15 per hour and the labelers were informed of the task and compensation before completing the task. The complete set of labels are included in the dataset.

Task Definition for NICE-Setting I. We define NICE-Setting I as generating dialogue-styled comments for an image. Formally, the generation task as follows: given an image $\boldsymbol{I}_{image}$, and N comments $\boldsymbol{C}_1, ..., \boldsymbol{C}_N$. Systems aim to generate the comment $\boldsymbol{C}_k$, where k is from 1 to N using the current state information $S_{I_{image}, C_1, ..., C_{k-1}}$. The state information contains input image feature $\boldsymbol{I}_{image}$ and the comments history $(\boldsymbol{C}_1, ..., \boldsymbol{C}_{k-1})$.

4.2 NICE-Setting II (Auto Labeling)

NICE-Setting I provides human labels for a subset of the NICE dataset. We have annotated over 28,000 human samples. However, human labeling for the full dataset would be too onerous in terms of worker and financial resources since we have 2M images and the corresponding comments in the whole NICE dataset. To address this issue we use a weakly-supervised approach and generate affect features as a substitute. This forms the second setting for our analyses.

Auto Labeling for NICE-Setting II. In this task, we generate style and affect features for all the comments to facilitate controlling comment generation. The input in this case is a tuple that contains an image, the thread title, the current comment history, and affective feature for the targeted comment. We applied similar filters as in NICE-Setting I on the image and text and we treat the title of the thread as the “comment topic”. To further clean the data we remove some threads without any comments except the thread title and only keep the first five comments for each thread. After the cleaning, the dataset for this setting 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 [2].

Affect Features for Auto Labeling on NICE-Setting II. 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 use Linguistic Inquiry and Word Count (LIWC)(Pennebaker et al., 2001). LIWC is a tool which is widely used for text analysis in linguistics and psychology, and has been demonstrated to capture important information (Chung and Pennebaker, 2018). In this second setting, 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 normalized feature vector for

[2] Each image can have multiple dialogue threads.

each comment automatically by counting the number of words for each dictionary categories. We hypothesize that these features can represent the open-domain human affect and language style in the comments.

Task Definition for NICE-Setting II. We define the NICE-Setting II task as generating comments in response to an 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 generate the comment $\boldsymbol{C}_k$ using the current state information $S_{I_{image},H,C_1,...,C_{k-1}|A_k}$, which contains the input image features $\boldsymbol{I}_{image}$, comment topic $\boldsymbol{H}$, and the comments history $(\boldsymbol{C}_1, ..., \boldsymbol{C}_{k-1})$, and is conditioned on the affect feature $\boldsymbol{A}_k$.

5 Experiments

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

Baseline Models. We provide the results using "off-the-shelf" baseline models on the proposed NICE-Setting I to benchmark performance. This is important to provide a comprehensive picture of the current performance of state-of-the-art methods on this task. The details of the baseline models can be found in the Appendix C.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 2. The results show 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 labelers rated each generated comment in terms of how engaging, emotional, empathetic, appropriate and relevant it was. Table 2 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. However, based on the results, these models are far from capturing 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. One of the reasons is that the image captioning model generate descriptions based on specific objects in the image, while emotional content is more nature and more abstract. Performing perfectly on all criteria is challenging but we believe these systems can be improved to have better results. The qualitative examples for baseline models on NICE-Setting I are presented in Appendix C.2.

5.2 Experiments on NICE-Setting II

Pre-training model of MAGIC. Next, we present a novel large-scale pre-training model on the NICE-Setting II dataset. Our model (Modeling Affect Generation for Image Commenting, or MAGIC) aims to generate emotional comments conditioned on an image, a comment topic, affect features, and the comment history. We introduce the MAGIC model and our training procedure in the following section.

MAGIC Training. As large models usually generalize better to new domains when they are trained on large volumes of data, we use GPT-2 (Radford et al., 2019) as the backbone for MAGIC. It is trained with the objective of predicting the next word, given an image, comment topic, comment history, affect feature, and all of the previous words within a defined context window. We trained MAGIC with the transformer architecture, which

	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 2: Performance on the NICE-Setting I. 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).

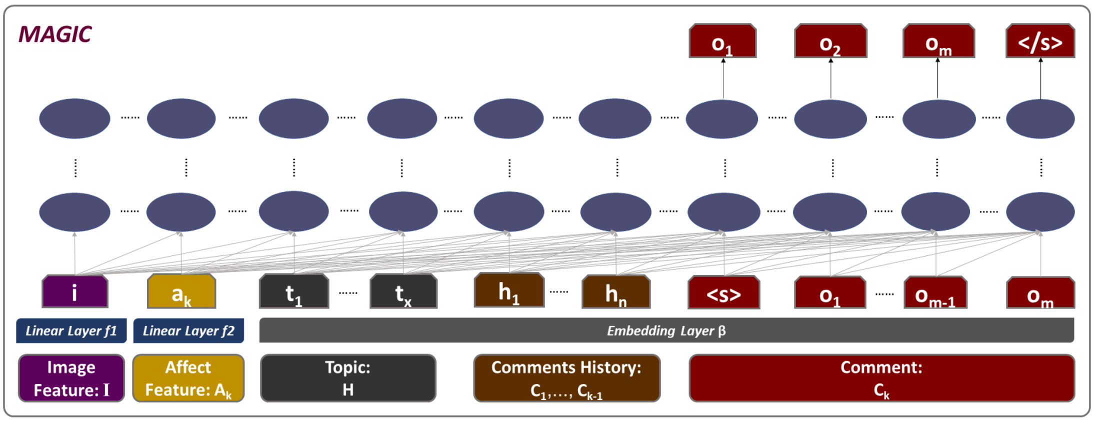


Figure 5: An overview of our MAGIC model.

has 12 layers and each layer has 12 heads. Based on the definition in 4.2, the model aims to compute the conditional probability L:

$$L = P(\boldsymbol{C}_k | \boldsymbol{I}_{image}, \boldsymbol{H}, \boldsymbol{A}_k, \boldsymbol{C}_1, ..., \boldsymbol{C}_{k-1}) \qquad (1)$$

When training MAGIC, as shown in Fig. 5, we encode the input image $\boldsymbol{I}_{image}$ into a 2048-dimension feature vector using pre-trained Resnet-152 model (He et al., 2016). The affect and style features $\boldsymbol{A}_k$ (introduced in 4.2) are represented as a 64-dimensional affect feature vector. The image feature vector and affect feature vector are passed to two separate linear layers to map to two 768-dimension vectors $\boldsymbol{i}$ and $\boldsymbol{a}_k$. Then, the comment topic $\boldsymbol{H}$, comment history $(\boldsymbol{C}_1, ..., \boldsymbol{C}_{k-1})$ and output comments $\boldsymbol{C}_k$ are fed into an embedding layer β to generate embedding vectors for each set of tokens respectively, $\boldsymbol{t}_1, ..., \boldsymbol{t}_x$, $\boldsymbol{h}_1, ..., \boldsymbol{h}_n$ and $\boldsymbol{o}_1, ..., \boldsymbol{o}_m$ for each token as follows:

$$E_{topic} = \boldsymbol{t}_1, ..., \boldsymbol{t}_x = \beta(\boldsymbol{H}) \qquad (2)$$

$$E^k_{history} = \boldsymbol{h}_1, ..., \boldsymbol{h}_n = \beta(\boldsymbol{C}_1, ..., \boldsymbol{C}_{k-1}) \qquad (3)$$

$$E^k_{comment} = \boldsymbol{o}_1, ..., \boldsymbol{o}_m = \beta(\boldsymbol{C}_k) \qquad (4)$$

The encoded image feature vector $\boldsymbol{i}$, the affect feature vector $\boldsymbol{a}_k$, 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 follows:

$$\boldsymbol{B}^k = f_{concat}(\boldsymbol{i}, \boldsymbol{a}_k, E_{topic}, E^k_{history}, E^k_{comment}) \qquad (5)$$

Then, $\boldsymbol{B}^k$ is fed to the 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:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{M} \circ \boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V} \qquad (6)$$

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

Inference and Learning Strategy of MAGIC. Given a training dataset with D samples, all comments in each sample has a total of Y tokens. We maximize the log-likelihood (MLE) to learn the model parameters θ of the conditional probabilities

	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 3: Results of four models on the 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.

L_θ over the entire training dataset:

$$B^{k,m} = f_{concat}(\boldsymbol{i}, \boldsymbol{a}^k, E_{topic}, E^k_{history}, \boldsymbol{o}^k_1, ..., \boldsymbol{o}^k_m) \quad (7)$$

$$L_\theta(D) = \sum_{i=1}^{D} \sum_{m=1}^{Y} p_\theta(\boldsymbol{o}^k_m | B^{k,m-1}) \quad (8)$$

During inference, each token is generated one by one via beam search with a beam size of two.

Implementation of MAGIC. We split the 6,720,536 image-comment pairs of NICE-Setting II data to 6,550,536 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 32 on each GPU using a machine with 4xV100 32G GPUs and the learning rate was $5e-5$. Total training time is about 7 days.

Evaluation of MAGIC. 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), and compared them with our MAGIC model. Details about how we modified the baseline models are described in Appendix D.1. Table 3 shows the performance of our MAGIC model and baseline 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, 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) to evaluate 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 BUTD model. More samples of Generated Image Comments by MAGIC on NICE-Setting II are included in Appendix D.2.

5.3 Adapt Pre-training MAGIC Model to Domain-Specific Task

Pre-training MAGIC model is also flexible to be adapted to related domain-specific conditional generation tasks. The affect feature can be replaced with emotional or personalized features for other conditional image-text generation tasks. Appendix E show an experiment that we adapt trained MAGIC model to Personality-Captions dataset (PCD) (Shuster et al., 2019). The result show that our MAGIC pre-training model has good performance on another similar task (PCD).

6 Conclusion and Future Work

In this paper we present a new vision-language task called Neural Image Commenting with Empathy (NICE) which extends image descriptions to comments with an emphasis on emotion and affect. We design contexts for this dataset based on different annotation schemes. For NICE-Setting II, we propose a novel per-trained model, MAGIC, for image commenting conditioned on affect and style features. We show that MAGIC can help produce affective and emotional image comments. To facilitate research in this area, we release the NICE dataset. The social language captured in this dataset has great value for training conversational systems. Image commenting is an emerging area of research and AI systems for conversation are

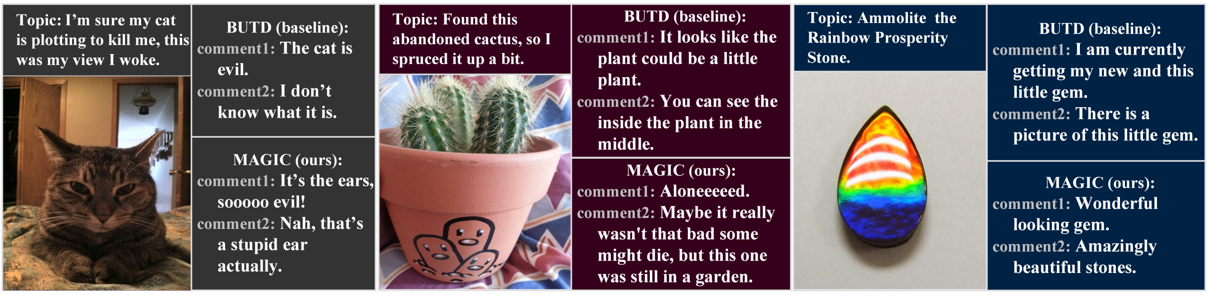

Topic: I'm sure my cat is plotting to kill me, this was my view I woke.
BUTD (baseline):
comment1: The cat is evil.
comment2: I don't know what it is.
MAGIC (ours):
comment1: It's the ears, sooooo evil!
comment2: Nah, that's a stupid ear actually.
Topic: Found this abandoned cactus, so I spruced it up a bit.
BUTD (baseline):
comment1: It looks like the plant could be a little plant.
comment2: You can see the inside the plant in the middle.
MAGIC (ours):
comment1: Aloneeeeed.
comment2: Maybe it really wasn't that bad some might die, but this one was still in a garden.
Topic: Ammolite the Rainbow Prosperity Stone.
BUTD (baseline):
comment1: I am currently getting my new and this little gem.
comment2: There is a picture of this little gem.
MAGIC (ours):
comment1: Wonderful looking gem.
comment2: Amazingly beautiful stones.

Figure 6: Examples of comments generated using the MAGIC model on NICE-Setting II.

becoming increasingly widely adopted. 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.

Broader Impact

Visual text generation has many applications. In addition to commenting, grounded language models could help drive content generation for bots and AI agents, and assist in productivity applications, helping to re-write, paraphrase, translate or synthesize text. Fundamental advances in text generation help contribute towards these goals and many would benefit from a greater understanding of how to model emotional and empathetic language. Arguably many of these applications could have positive benefits. However, this technology could also be used by bad actors. AI systems that generate content can be used to manipulate or deceive people. Therefore, it is very important that this technology is developed in accordance with responsible AI guidelines. For example, explicitly communicating to users that content is generated by an AI system and providing the user with controls in order to customize such a system. It is possible our dataset could be used to develop new methods to detect manipulative content - partly because it is rich with emotional language -and thus help address another real world problem.

Our dataset is collected from the licensed website, which is not a fully representative source. Therefore, we also need to understand biases that might exist in this corpus. Data distributions can be characterized in many ways. The release of this dataset will be done in accordance with copyright law. We will release links to content that is already in the public domain. Moreover, we have filtered sensitive content, which helps reduce the risk of harmful content within the dataset. Thus, it may not be considered a fully representative source. In this paper, we have captured how the word level distribution in our dataset is different from other existing datasets. However, there is much more than could be included in a single paper. We would argue that there is a need for more datasets linked to real world tasks and that by making these data available we can help researchers answer these questions.

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.

Appendix

A Details of Filter for NICE Dataset

It took several researchers multiple weeks to remove sensitive content for both image and text filtering. For example, we used the “Microsoft Adult Filtering API” (Microsoft, 2019) to remove adult, racy and gory images, we use the “Detecting image types API” (Microsoft, 2018) to remove clip art and line drawings, we use the “Optical Character Recognition (OCR) API” (Microsoft, 2020) to remove printed or handwritten text from the images, such as photos of license plates or containers with serial numbers, as well as from documents invoices, bills, financial reports, articles, and more. We also removed people’s names, politically sensitive language, ethnic-religious content, or other potentially offensive material (including inappropriate references to violence, crime and illegal substances) as the similar filter API for language cleaning.

The last step of filtering, we make sure that NICE dataset had no more than 5 (≤ 5) corresponding comments for each image, and there are not more than 6 (≤ 6) different dialogue threads for the same image.

In NICE-Setting II, after annotation, we filter out image-comments pairs without affect feature or dialogue topic from dialogue thread. We will keep cleaning and maintaining it in future.

B Human Labeling Task for NICE-Setting I

Screenshot of the human labeling. A screenshot of the human labeling task on M-Turk is shown in Fig. 7.

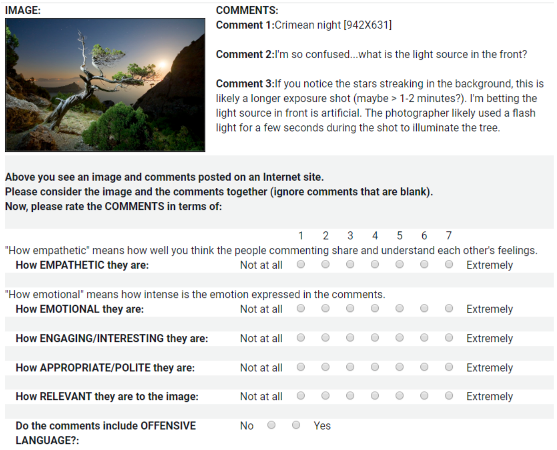


IMAGE:

COMMENTS:
Comment 1: Crimean night [942X631]

Comment 2:I'm so confused...what is the light source in the front?

Comment 3: If you notice the stars streaking in the background, this is likely a longer exposure shot (maybe > 1-2 minutes?). I'm betting the light source in front is artificial. The photographer likely used a flash light for a few seconds during the shot to illuminate the tree.

Above you see an image and comments posted on an Internet site.
Please consider the image and the comments together (ignore comments that are blank).
Now, please rate the COMMENTS in terms of:

"How empathetic" means how well you think the people commenting share and understand each other's feelings.

"How emotional" means how intense is the emotion expressed in the comments.

		1	2	3	4	5	6	7	
How EMPATHETIC they are:	Not at all	○	○	○	○	○	○	○	Extremely
How EMOTIONAL they are:	Not at all	○	○	○	○	○	○	○	Extremely
How ENGAGING/INTERESTING they are:	Not at all	○	○	○	○	○	○	○	Extremely
How APPROPRIATE/POLITE they are:	Not at all	○	○	○	○	○	○	○	Extremely
How RELEVANT they are to the image:	Not at all	○	○	○	○	○	○	○	Extremely

Do the comments include OFFENSIVE LANGUAGE?: No ○ ○ Yes

Figure 7: A screenshot of the human labeling task on NICE-Setting I.

Heuristic for filtering. We also created a heuristic for filtering “good” comments from “bad”. The comment had to satisfy the following criteria:

$$\begin{aligned} \text{Appr.} > 1 \text{ AND Emp.} > 1 \text{ AND Relev.} > 1 \text{ AND} \\ \mu(\text{Appr.,Emp.,Emotion,Relevance}) > 3 \text{ AND} \\ \text{Offensive} == \text{No} \end{aligned} \tag{9}$$

Of the 28,392 images 20,000 (70%) satisfied this criteria. These filtered image constitute a “clean" set of data.

C Appendix for NICE-Setting I

C.1 Baseline Models on NICE-Setting I

Vision-Language Pre-Training (VLP). Large-scale language pretrained models relying on massive data and self-supervised learning tasks like masking have created a new state-of-the-art in several natural language processing tasks (Devlin et al., 2018). Pretraining models across language and vision poses a challenging task where usually the amount of training data is several times smaller than the text only pretraining. Among various vision-language pretraining models proposed recently (Sun et al., 2019; Li et al., 2019, 2020; Su et al., 2020), and one of them (Zhou et al., 2019) performed both classification (e.g., VQA) and generation (image captioning). To use VLP (Zhou et al., 2019), we pretrain the model on the large scale Conceptual Captions dataset (Sharma et al., 2018b) that consists of 3 million image-text pairs. We then fine tune the pre-trainied model on the NICE-Setting I dataset with captioning loss only (minimizing perplexity) and report the results.

Bottom-UP Top-Down Attention (BUTD). Using pretrained object detectors for image captioning has resulted in significant performance gains compared to using CNN features as shown in Anderson et al. (2018). We use this model as a baseline on the NICE-Setting I dataset.

Semantic Compositional Networks (SCN). SCNs (Gan et al., 2017) rely on a pretrained tagger to provide visual cues about the entities and actions in an image, and leverage LSTMs to generate a natural language description for images. Using this model can also help us to understand the performance difference between a tagger based model (SCN) and an object detection based model (BUTD and VLP).

Microsoft Captioning System (Caption-Bot). The Microsoft image captioning bot (Microsoft, 2017) is a publicly available agent that can generate descriptions for a given image.

LSTM based caption generation (LSTM-XE). LSTM based image captioning (Vinyals et al., 2015) was one of the first models proposed to use pretrained CNNs as in conjunction with an LSTM based language model, which to generate descriptions for images. It is our final baseline on NICE-Setting I.

C.2 Qualitative Examples for baseline models on NICE-Setting I

Fig. 8 shows examples of comments generated by each baseline model for three images on NICE-Setting I. We observe the comments generated by baseline models are reasonable in content but not very emotional, subjective or imaginative in the context of social dialogue, and thus less likely to lead to user engagement. We hope that the benchmark baselines provided will serve as a reference for researchers, and inspire the creation of more appropriate models for human-machine interaction on NICE dataset.

D Appendix for NICE-Setting II

D.1 Implementation Details of Experiment for MAGIC on NICE-Setting II

In this section, we describe the implementation of our baselines in the experiments. We modified Show Attention and Tell (ShowAttTell) (Xu et al., 2015) and Bottom-Up-Top-Down Attention (BUTD) (Anderson et al., 2018) models to the image commenting task. In this task, the inputs are tuples of the image, the affect feature, a mood topic and the comment history, and the output is a comment. For both models, we use a linear layer to map the 64-dimension affect feature to 512 dimensions. The mood topic is concatenated with the comment history and passed to an embedding layer.

In ShowAttTell, the decoder computes a weighted image attention vector at each time step, and uses it to generate a text token. To adapt this model on image commenting task, we concatenate the weighted image attention vector with the 512-dimension affect vector, the embedded topic and the comment history. This new concatenation vector replaces the original image attention vector and is used to generate the comment token at each time step.

In BUTD decoder, a top-down attention module computes an attention vector on image and passes it to a language module. The language module takes the image attention vector to generate text token at each time step. We use the similar modification that the concatenation of the image attention vector, the 512-dimension affect vector, the embedded topic and the comment history, which is passed to language model for comment decoding. In both models, the embedding size is 512 dimensions, the hidden size of LSTMs is 1024 demensions and they are trained by optimizing the cross-entropy loss with a learning rate 5e-4.

For the ablation study, we use the GPT-2 (Radford et al., 2019) trained on NICE dataset without affect vector (LIWC feature). Thus, the input for GPT-2 only has the mapped the image features, the embedded mood topic and the comment history. By optimizing the cross-entropy loss, GPT-2 is trained 30 epochs on NICE dataset.

D.2 Samples of Generated Image Comments by MAGIC on NICE-Setting II

In Figure 9, we show some samples generated from MAGIC model on test set of NICE-Setting II dataset. Each example contains an image, a topic which is the thread title of a dialogue post, and the generated comments.

E Adapt MAGIC to Domain-Specific Tasks

E.1 Adapt MAGIC to PCD

The pre-training MAGIC model is flexible to be adapted to related domain-specific conditional generation tasks. The affect feature can be replaced with emotional or personalized features for other conditional image-text generation tasks. For example, Personality-Captions dataset (PCD) (Shuster et al., 2019) defined 215 categories of personality traits. Based on the hypothesis that the affect feature can model general language affect or styles, MAGIC pre-training model learns the relations between image, comment topic, affect feature and comments. To adapt trained MAGIC on PCD, the personalities are embedded via an embedding layer to 64-dimension vectors. As same dimension as affect features on NICE, these vectors replace affect feature as input in MAGIC on PCD dataset. We fine-tuning MAGIC on PCD to generate captions which are conditioned on pre-defined 215 domain-specific personalities.

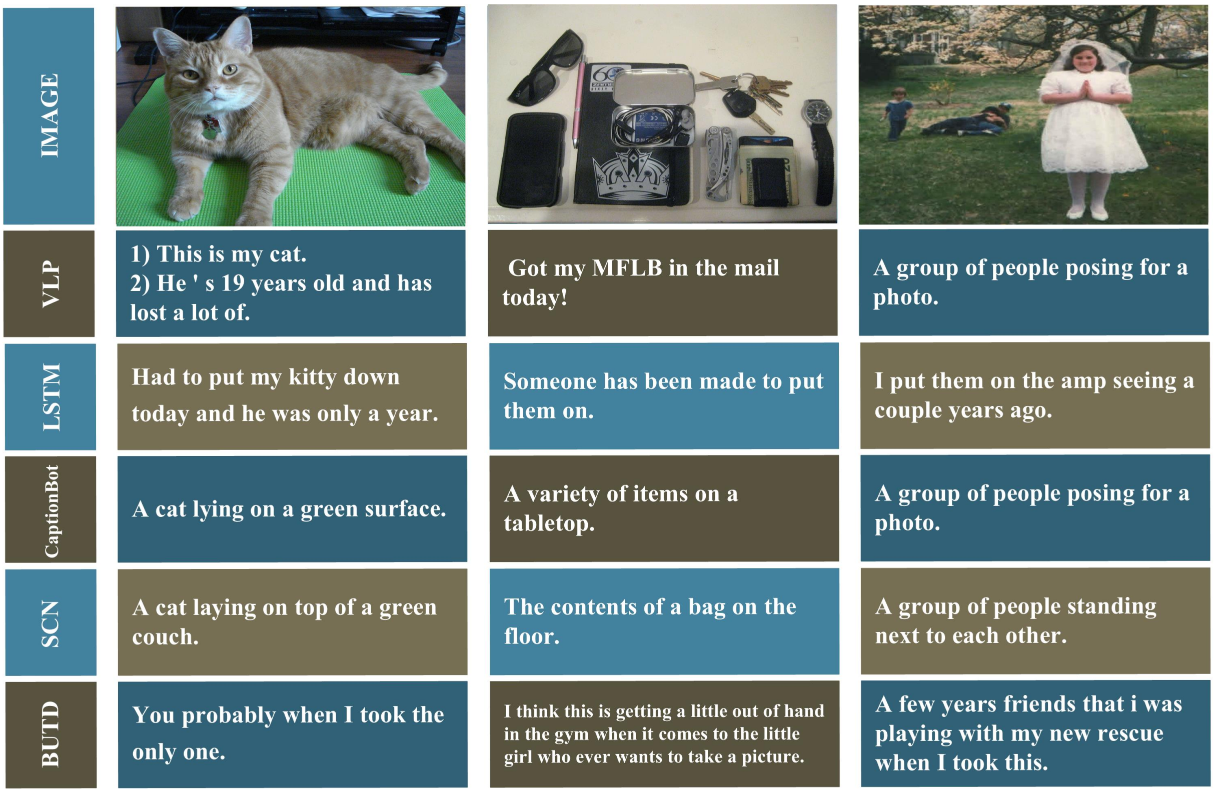
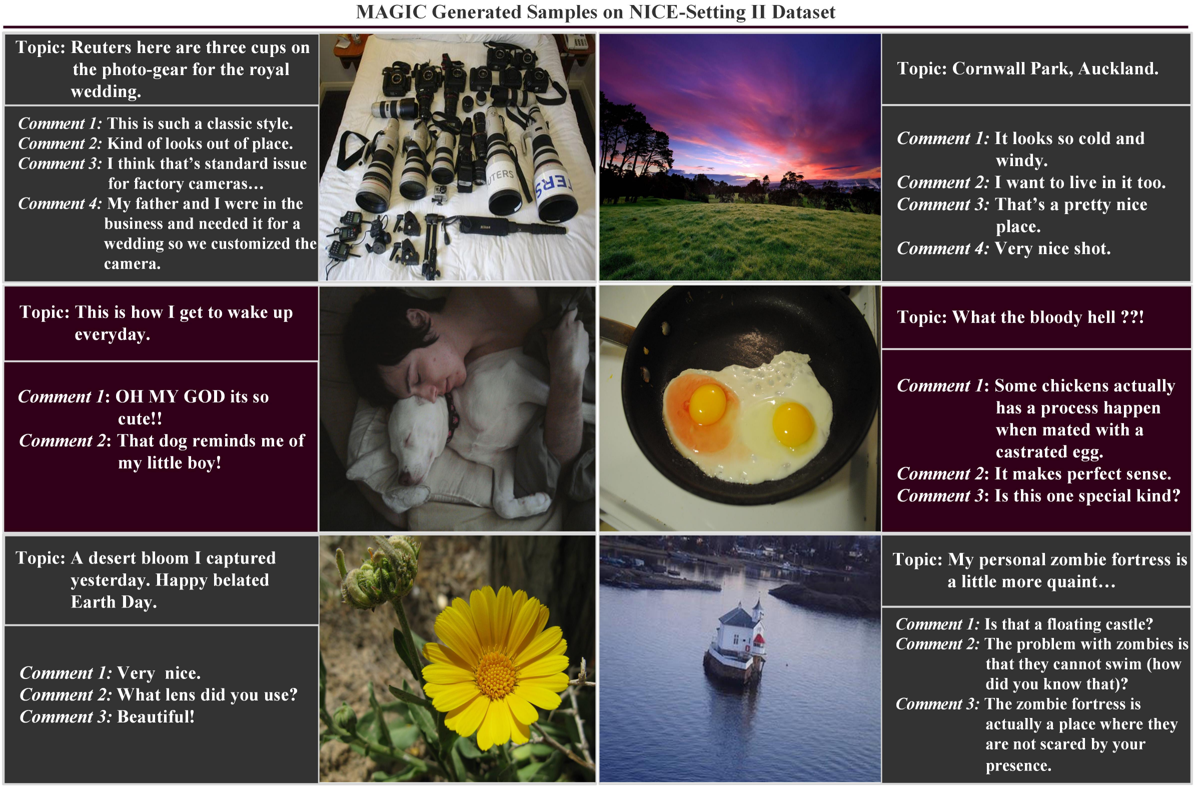


IMAGE			
VLP	1) This is my cat. 2) He ' s 19 years old and has lost a lot of.	Got my MFLB in the mail today!	A group of people posing for a photo.
LSTM	Had to put my kitty down today and he was only a year.	Someone has been made to put them on.	I put them on the amp seeing a couple years ago.
CaptionBot	A cat lying on a green surface.	A variety of items on a tabletop.	A group of people posing for a photo.
SCN	A cat laying on top of a green couch.	The contents of a bag on the floor.	A group of people standing next to each other.
BUTD	You probably when I took the only one.	I think this is getting a little out of hand in the gym when it comes to the little girl who ever wants to take a picture.	A few years friends that i was playing with my new rescue when I took this.

Figure 8: Example comments to user-shared images generated by the baseline models on NICE-Setting I.



MAGIC Generated Samples on NICE-Setting II Dataset

Topic: Reuters here are three cups on the photo-gear for the royal wedding.

Comment 1: This is such a classic style.
Comment 2: Kind of looks out of place.
Comment 3: I think that's standard issue for factory cameras…
Comment 4: My father and I were in the business and needed it for a wedding so we customized the camera.

Topic: Cornwall Park, Auckland.

Comment 1: It looks so cold and windy.
Comment 2: I want to live in it too.
Comment 3: That's a pretty nice place.
Comment 4: Very nice shot.

Topic: This is how I get to wake up everyday.

Comment 1: OH MY GOD its so cute!!
Comment 2: That dog reminds me of my little boy!

Topic: What the bloody hell ??!

Comment 1: Some chickens actually has a process happen when mated with a castrated egg.
Comment 2: It makes perfect sense.
Comment 3: Is this one special kind?

Topic: A desert bloom I captured yesterday. Happy belated Earth Day.

Comment 1: Very nice.
Comment 2: What lens did you use?
Comment 3: Beautiful!

Topic: My personal zombie fortress is a little more quaint…

Comment 1: Is that a floating castle?
Comment 2: The problem with zombies is that they cannot swim (how did you know that)?
Comment 3: The zombie fortress is actually a place where they are not scared by your presence.

Figure 9: Generated Samples from MAGIC Model on NICE-Setting II Dataset.

E.2 Analysis of Adapting MAGIC to PCD

	Embedding Similarity				Diversity	
Setting	**Spice**	**BertP**	**BertR**	**BertF1**	**Entro.4**	**Disti.2**
GPT2	0.032	0.244	0.286	0.252	11.110	0.399
GPT2-NoAffect	0.032	0.246	0.286	0.254	11.073	**0.408**
MAGIC	**0.035**	**0.248**	**0.291**	**0.257**	**11.145**	0.399

Table 4: Automatic Evaluation results of on Personality-Captions dataset. Comparing with two baselines using GPT2 (Radford et al., 2019), MAGIC has good transfer learning ability on domain-specific tasks.

PCD contains (image, personality trait, caption) triples collected using crowd-workers and has train/val/test splits with 186,858/5,000/50,000 samples. We use the adaption method to train MAGIC continually. In PCD dataset, each image only has one corresponding caption, there aren't any comment history and comment topic. We evaluate three different models. The first one is a GPT-2 model without training on NICE dataset; the second is GPT-2 trained on NICE dataset without affect features (GPT-2-NoAffect); and the last is the standard MAGIC model trained on NICE. For each model, to test the transfer learning ability of MAGIC, we trained 20 epochs on PCD dataset. As the personality traits have less information than affect feature, and generated utterances from pre-trained MAGIC have high variety, token matching metrics are not appropriate to evaluate the performance. Only embedding similarity metrics and diversity metrics are showed in Table 4.

From the embedding similarity results, MAGIC performs better than the other two models. This demonstrates that MAGIC has better transfer learning ability for the similar domain-specific tasks involving affect or personalities. From the diversity metrics, three models are close with each other. One main reason is that all three models are pre-trained on a large number of data, which provides a high diversity language patterns, which allows for more human-like outputs. Appendix E.3 contains generated samples on PCD.

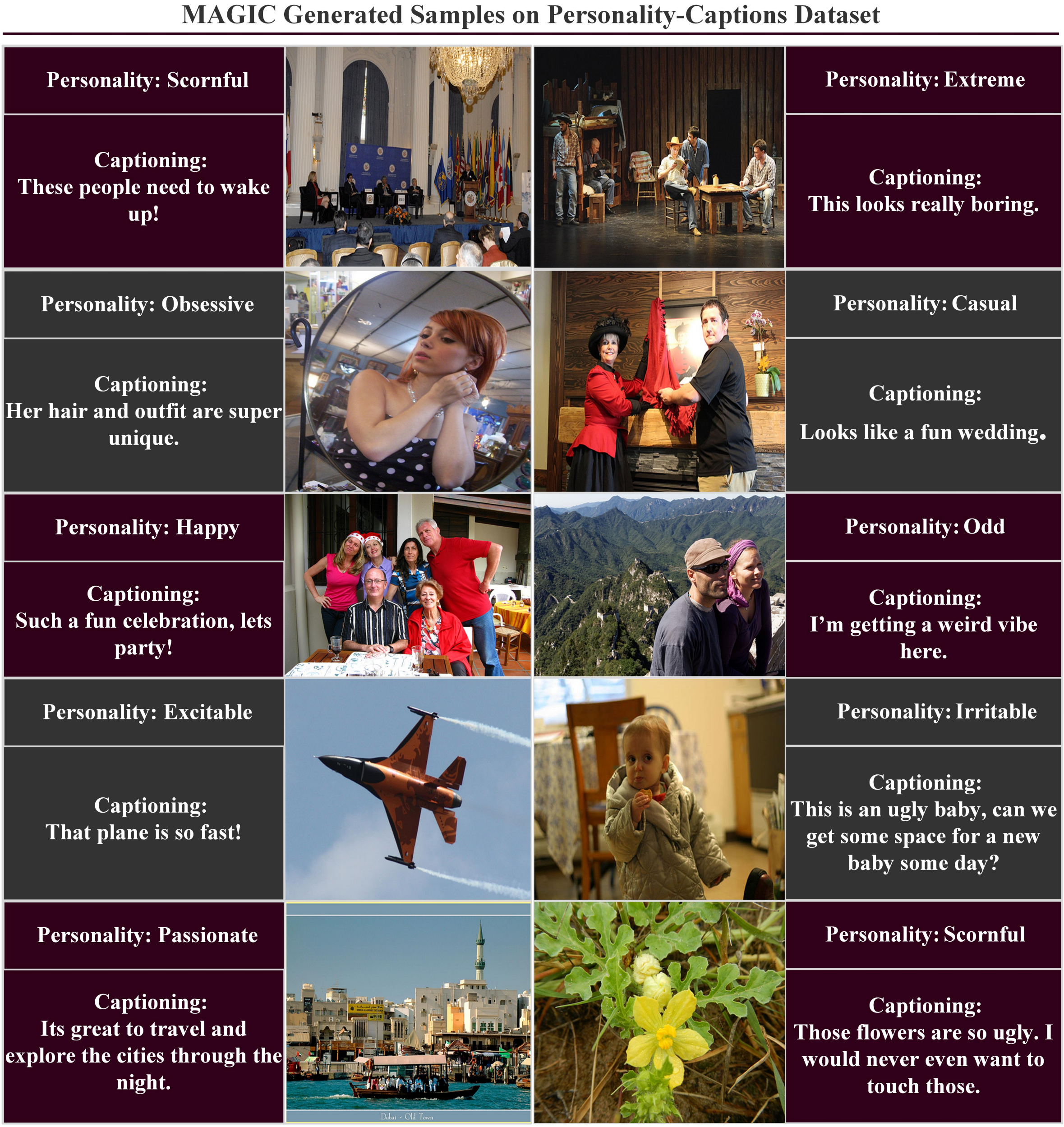
Human Evaluation. We perform six human evaluation tasks using Amazon Mechanical Turk: Personality, Appropriate, Emotional, Empathetic, Engaging, and Relevant. For each task, we use 500 test image sequences from Personality-Captions dataset. We compare the MAGIC model with GPT-2 model. During human evaluation, each rater was displayed the images, personality traits and the generated captions. The raters were asked to rate from a 7 point scale on six different aspects: how the generated caption matched the personality, whether it was appropriate, emotional, empathetic, engaging and relevant to the images. The results of the average scores on each aspects across 500 samples are in table 5. From the results, MAGIC outperforms the GPT-2 on all aspects. The personality, emotional and engaging have more significant difference than the other three aspects. This indicates that adapting MAGIC on Personality-Captions dataset can generate more human-like comments than GPT-2.

Model	**Perso.**	**Appro.**	**Emo.**	**Empath.**	**Engag.**	**Relev.**
GPT-2	3.55	4.52	3.72	3.81	4.34	4.71
MAGIC	**3.68**	**4.58**	**4.04**	**3.86**	**4.45**	**4.78**

Table 5: Human evaluation results.

E.3 Samples of Generated Captions from MAGIC on Personality-Captions Dataset

In Figure 10, we show some samples generated by adapting MAGIC model on Personality-Captions dataset. Each example contains an image, a personality, and the generated comments.



MAGIC Generated Samples on Personality-Captions Dataset

Personality: Scornful

Captioning:
These people need to wake up!

Personality: Extreme

Captioning:
This looks really boring.

Personality: Obsessive

Captioning:
Her hair and outfit are super unique.

Personality: Casual

Captioning:
Looks like a fun wedding.

Personality: Happy

Captioning:
Such a fun celebration, lets party!

Personality: Odd

Captioning:
I'm getting a weird vibe here.

Personality: Excitable

Captioning:
That plane is so fast!

Personality: Irritable

Captioning:
This is an ugly baby, can we get some space for a new baby some day?

Personality: Passionate

Captioning:
Its great to travel and explore the cities through the night.

Personality: Scornful

Captioning:
Those flowers are so ugly. I would never even want to touch those.

Figure 10: Generated Samples from MAGIC Model on Personality-Captions Dataset.

References

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. *Proceedings of EMNLP*.

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. *Proceedings of ECCV*.

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. *Proceedings of NAACL*.

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. *Proceedings of ACL*.

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.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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. *Proceedings of ICML*.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of AAAI*.

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

Chin-Yew Lin. 2004. Rouge: A package for 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. *Proceedings of ECCV*.

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. *Proceedings of NeurIPS*.

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. 2014. Personal productivity assistant in microsoft 365. https://www.microsoft.com/en-us/cortana.

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

Microsoft. 2018. Detecting image types api. https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-detecting-image-types.

Microsoft. 2019. Microsoft adult filtering api. https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-detecting-adult-content.

Microsoft. 2020. Microsoft optical character recognition api. https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. *Proceedings of ACL*.

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. *Proceedings of NAACL*.

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. *Proceedings of ACL*.

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. 2018a. 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*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018b. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

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.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. *Proceedings of CVPR*.

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

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *Proceedings of EMNLP*.

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

Lucy Vanderwende, Arul Menezes, , and Chris Quirk. 2015. An amr parser for english, french, german, spanish and japanese and a new amr-annotated corpus. *NAACL*.

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.

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. *Proceedings of ICML*.

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. *Proceedings of ICLR*.

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. *Proceedings of AAAI*.