Appendix:
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

1 Sentiment Words in NICE dataset

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

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.

2 Appendix for NICE-Setting I

2.1 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., 2019b,a; Su et al., 2019), 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., 2018) 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.

[*] Equal Contribution.

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

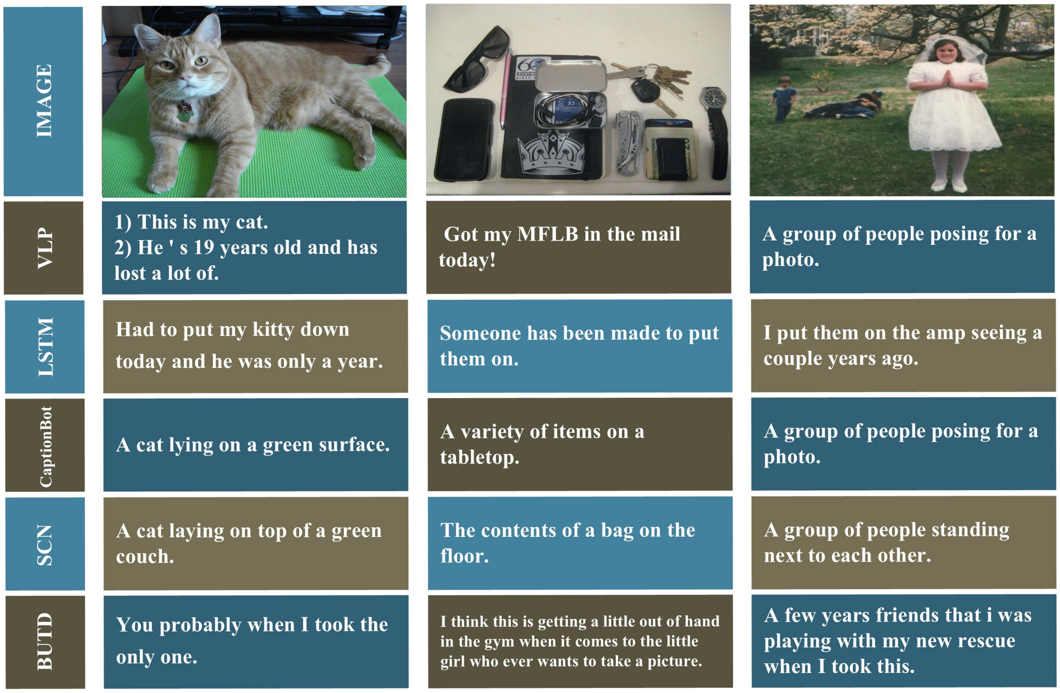


IMAGE			
VLP	1) This is my cat. 2) He ' s 19 years old and has lost a lot of.	Got my MFLB in the mail today!	A group of people posing for a photo.
LSTM	Had to put my kitty down today and he was only a year.	Someone has been made to put them on.	I put them on the amp seeing a couple years ago.
CaptionBot	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 1: Example comments to user-shared images generated by the baseline models on NICE-Setting I.

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.

2.2 Qualitative Examples for baseline models on NICE-Setting I

Fig. 1 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.

3 Appendix for NICE-Setting II

3.1 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.

3.2 Samples of Generated Image Comments by MAGIC on NICE-Setting II

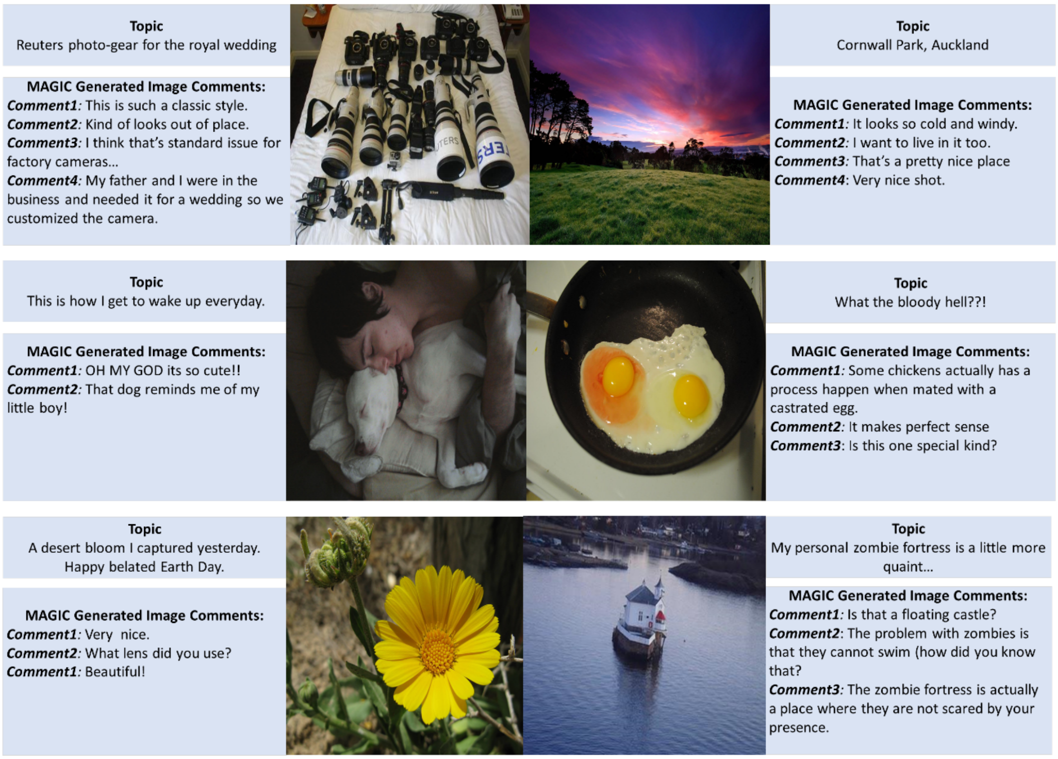
In Figure 2, we show some samples generated from MAGIC model on test set of NICE-Setting II dataset. Each example contains an image, a topic which is the thread title of a dialogue post, and the generated comments.

4 Impact Statement

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 Internet, 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. 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.



Topic
Reuters photo-gear for the royal wedding

MAGIC Generated Image Comments:
Comment1: This is such a classic style.
Comment2: Kind of looks out of place.
Comment3: I think that's standard issue for factory cameras...
Comment4: My father and I were in the business and needed it for a wedding so we customized the camera.

Topic
Cornwall Park, Auckland

MAGIC Generated Image Comments:
Comment1: It looks so cold and windy.
Comment2: I want to live in it too.
Comment3: That's a pretty nice place
Comment4: Very nice shot.

Topic
This is how I get to wake up everyday.

MAGIC Generated Image Comments:
Comment1: OH MY GOD its so cute!!
Comment2: That dog reminds me of my little boy!

Topic
What the bloody hell??!

MAGIC Generated Image Comments:
Comment1: Some chickens actually has a process happen when mated with a castrated egg.
Comment2: It makes perfect sense
Comment3: Is this one special kind?

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

MAGIC Generated Image Comments:
Comment1: Very nice.
Comment2: What lens did you use?
Comment1: Beautiful!

Topic
My personal zombie fortress is a little more quaint...

MAGIC Generated Image Comments:
Comment1: Is that a floating castle?
Comment2: The problem with zombies is that they cannot swim (how did you know that?
Comment3: The zombie fortress is actually a place where they are not scared by your presence.

Figure 2: Generated samples from MAGIC model on NICE-Setting II Dataset.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

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*.

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*.

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*.

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*.

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

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

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

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

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. *arXiv preprint arXiv:1502.03044*.

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*.