How do people talk about images?
A study on open-domain conversation on images.

Anonymous ACL submission

Abstract

Open-domain conversation on images requires the model to consider the relation and balance between utterances and images in order to generate proper responses. This paper explore how human conduct conversation on images by investigating a well-constructed open-domain image conversation dataset, ImageChat. We examine the conversations on images from three perspectives: *image relevancy*, *image information* and *utterance style*. We show that objects in the image are indeed the most important element for conversations on image, which could be directly discussed or be a bait to other off-image conversations. Thus, being able to accurately detect objects in the image and knowing their attributes are essential to chat on image. Understanding the scenarios of the image, except extracting the image objects, is also a key factor to the conversation on images. Based on our analysis, we propose to enriching the image information with image caption and object tags, increasing the diversity and image-relevancy of generated responses. We believe that our analysis provides useful insights and directions that facilitate future research on open-domain conversation on images.

1 Introduction

A picture is worth a thousand words. Human communication often involves multi-media including text and image. Understanding the image content and chat about it is an important skill for a chatbot to interact with people. Despite of the flourish research on open-domain dialogue systems, most of them focus on text-based conversation without multimedia inputs (Zhang et al., 2020; Adiwardana et al., 2020; Roller et al., 2021). Current multimodal dialogue systems often adapt similar Transformer-based architecture as text-based dialogue systems, and focus on fusing text and image modalities through complex attention mechanism (Ju et al., 2019) or simply concatenating image and text features and relying on multi-task training on multiple dialogue datasets (Shuster et al., 2020b,c).

The state-of-the-art model, Multimodal BlenderBot (MMB) (Shuster et al., 2020c), sometimes fails to chat about the images and generate general and non-informative response that is not related to the image. For example, MMB often generates “I think you are right" while it only appears once in the human generated references. Additionally, it sometimes changes the conversation topic abruptly when chatting in a long run. For example, given an image of cat, MMB first talks about cats and dogs, but then randomly switch to the topic about reading (Shuster et al., 2020c). These motivate us to investigate the conversation topics on images and its transition, and how the image features take parts in the conversation on images.

To understand what are the facts that direct the conversation on images and to know what types of image information is influencing the generation of image-related utterances, we investigate the ImageChat dataset (Shuster et al., 2020a) and conduct a deep analysis on the dataset. Thus, we ask the following questions: (1) Are the conversations in ImageChat always on the image-related theme? Or the image usually just serve as a bait and the conversations are focusing on other topics? What are the relation between those topics and the given image if in the latter case? (2) What types of the image information are used in the conversation? To be more specifically, how helpful the image objects are in the conversation on images? Since the baseline models usually use object detection model as the image encoder, we want to know what role the image objects play in the conversation on image. (3) How does the speaker’s style influence the conversation on images? The conversation in ImageChat was conducted when the speaking style of speakers is assigned, and there has been research on considering personality in dialogues. We are curious about how the speaking style affects the

conversation on images.

We found that the image objects and non-object image-related information like the image scenario are important in the image-related conversations (which are appeared in 55.9% and 29.4% of the utterances), and the scenario involving the existence of one of the image objects (even just a small part) often triggers the non-image-related conversations. Based on our analysis, we propose to enhance the image-dependent response generation by augmenting the image information from image caption and object tags, and using objects features rather than the whole image. Image scenario is critical for the conversation on images, and we try to get this information from the image caption. Our results using the enhanced image features outperforms the strong baseline model MMB, generating more image-related and diverse responses.

This is the first qualitative analysis for open-domain conversation on images, to the best of our knowledge. We analyze conversations on images to know what are the dominant factors in a conversation on image, and we propose several directions to aid the research in the open-domain conversation on image (See Sec. 2.3). Based on our findings, we improve the baseline model MMB with enhanced image features and generate more diverse and image-dependent responses.

2 Analysis of Conversations on Image

2.1 ImageChat Dataset

We analyze the ImageChat dataset (Shuster et al., 2020a) which is so far the only dialogue dataset that focuses on *open-domain conversation on images*, to the best of our knowledge. Each conversation is paired with one image from YFCC 100M (Thomee et al., 2016) and consisted of three turns utterances from two speakers with the assigned speaking styles. There are total 215 style types, such as sympathetic, optimistic, or dramatic, which are belong to one of the three categories: positive (81 styles), neutral (36 styles), and negative (98 styles). The images are highly diverse images across multiple domains. To understand more about the images, we use the Scene Graph Benchmark (Han et al., 2021) implementation of Faster R-CNN (Ren et al., 2016) to obtain the object tags from the image. Note that Faster R-CNN is also the image encoder used in the baseline system MMB. We also generate the caption of each image using the state-of-the-art language-vision pretrained model VinVL (Zhang et al., 2021). The statistics of the ImageChat dataset is shown in Table 1.

Category	Train	Valid	Test
images	186,782	4,999	9,997
dialogues	186,782	4,999	9,997
utterances	335,862	14,997	29,991

Table 1: Statistics of ImageChat dataset (Shuster et al., 2020a).

2.2 Aspects of Conversations on Image

We randomly sampled 108 utterances (36 conversations) from the validation set. Each conversation contains 1 image and 3 turns utterances of two speakers with different speaking styles. We annotate each utterance to find out the key factors for a open-domain dialogue on images by answering the following questions :

- Is the conversation theme always related to the image? If not, how does the conversation theme evolve?
- Do image objects help to reconstruct the dialogue utterances? What types of image information are helpful to reconstruct the conversation on images?
- Is the speaker's style an essential part to direct the conversation on images?

2.2.1 Image Relevance to Dialogue Theme

We first ask whether the conversation theme is always related to the image, and if not, how often is each utterance directly related to the image. We define the image relevancy as a binary classification of whether the given image is necessary for generating each utterance. If the image is referred in the utterance, the utterance is labeled as an image-related utterance; and if one could generate the utterance without the given image, the utterance is labeled as unrelated. Examples of image-related and unrelated utterances are shown in Table 2.

2.2.2 Image Information in the Dialogue

To know what kinds of image information is often mentioned in the dialogue on images, we labeled each utterance based on the type of image information. The baseline model rely on the object detection model to encode the image information. We assume the encoded features to be the objects

Image	Utterance	Related
	Cowardly: Never had this food before and not sure if I'm ready to try it today.	✔
	Appreciative (Grateful): I am always up to trying new things. It looks like a lot of effort went into this food and I plan to enjoy every bite.	✔
	Cowardly: I don't know, it looks like it might be too much.	✔
	Extraordinary: What an unusual place! The colors of the train really bounce off the grey backdrop of the city.	✔
	Narcissistic (Self-centered, Egotistical): Well, of course this is a fantastic picture, since it was MY magnificent photographic skills that produced it!	✔
	Extraordinary: I had no idea you have such talent!	✘
	Wise: Kids should have at least 2 hours of playtime per day	✘
	Reflective: However, I do think parents should have a say in what those kids are playing.	✘
	Wise: Parent involvement is always a good idea.	✘

Table 2: Examples of conversation themes are related and unrelated to the given image.

in the image, and we are curious about how helpful this objects would be for the multimodal dialogue system to reconstruct the dialogue utterances. Therefore, we first obtain the image object tags from the object detection model and investigate how frequent the object tags appears in the conversation. In addition, we also annotate whether there is other image-related information mentioned in the dialogue.

Base on our observation on the data, we compare the object tags with the utterance and categorize each utterance into one of the 8 classes indicating what type of the image information is mentioned in the utterance:

- T: There are words in the utterance exactly matching object tags.
- TN: There are words in the utterance referring to object tags, but not exactly match. For example, "guy" in the utterance, and "man" in the tag. Since the tags are high level labels, when there is hyponym in the utterance and hypernym in the tags, e.g. "seagull" in the utterance and "bird" in the tags, it also belongs to this category.
- TP: When pronoun is used in the utterance to refer to objects in the image, e.g. "what is she doing there?", "I can climb it".
- TF: When there are words in the utterance referring to objects in the image, but there is no match to the object tags. Probably because of the wrong object detection results.
- O: Other image-related information which is not object in the image is mentioned in the utterance.
- OS: The utterance is about the image itself, not objects or related objects in the image.
- OP: When pronoun is used in the utterance to refer to image-related information.
- N: There is no image-related information mentioned in the utterance.

See Table 3 for the example utterances of each category.

2.2.3 Style of the Utterance

We examine how much the utterance reflects the speaking style, after observing that some utterances are unnatural in the conversation. We score the utterance from 0 to 2 based on the degree of how much the utterance reflects the speaking style. 0 is

Class	Utterance	Image Tags
T	I guess this is an interesting **building**.	'cloud', 'window', 'sky', **'building'**
TN	I'd like to party with that **guy**!	'watch', 'glass', **'man'**, 'phone', 'hole', 'wall', 'guitar', ...
TP	Would **she** shut up already?	'book', 'microphone', 'jacket', 'tree', 'hair', **'woman'**, ...
TF	The **aluminum** art was different.	'rock', 'ground', 'foil'
O	It's obviously a **festival**.	'sunglasses', 'hat', 'speaker', 'light', 'balloon', 'balcony', ...
OS	A **screenshot** by definition does not die.	'man', 'hat', 'photo', 'glass'
OP	**It**'s beautiful! I would love to visit.	'leaf', 'flower', 'branch', 'tree'
N	yeah sure does.	'sunglasses', 'hat', 'man', 'ear', 'mouth', 'nose', 'light', ...

Table 3: Examples of each image information category of the utterance. The objects mentioned in the utterance (and in the image tags) are in bold.

given when utterance does not reflect the style, and 2 is given when the utterance is largely influenced by the speaking style. The examples are listed in Table 4.

2.3 Analysis Result and Finding

In this section, we describe the result and our analysis of the aforementioned questions. The three aspects of the conversation (image relevancy, image information, utterance style) are independent but intertwined. An utterance contains the exact image tag is not always image-related utterance (See analysis in Sec. 2.3.1), however, the result from image information analysis (Sec 2.3.2) suggests that image object tags (45.4%) and other image related information such as the scenario of the image (23.2%) take a large part in the image-related utterances.

To conclude, we point out directions to improve image-related utterance generation on the conversation on images based on our findings, including connecting objects to much broader scenarios, expanding the vocabulary size, obtaining the attributes of the objects, improving the object detection results, and using style control module in text generation models. More details analysis are described in the following subsections.

2.3.1 Image Relevancy

We find that conversation themes of ImageChat dialogues are not necessarily always about the image. In fact, the conversation often goes back and forth between image-related to non-related topics even within only three conversation turns. Figure 3 illustrates such phenomenon with dialogues of different combinations of the image-relevance utterances. While an image-related utterance is labeled as 'Y' and non-image-related utterance is labeled as 'N', 'YYY' means all three turns in a dialog are image-related utterances and 'YYN' means the conversation diverse from image-related topics to other domain not related to the given image.

Further investigating the combination of image-related and non-related utterances in a dialogue, we could roughly classify them into two schemas: (1) One speaker is responding the other, and if one extend out of the image-related topic, the following conversation diverse, vice versa. 'YNN', 'YYN', 'NYY' are in this category. The transition between 'Y' and 'N' sometimes is due to the mention of a related object, and sometimes people just invent a scenario in order to continue the conversation. (2) There are some dialogues seem unnatural because one of the speaker keep continuing his previous (self-)expression and not responding to the others utterance. 'YNY' and 'NYN' often belong to this schema. Note that there is no combination of 'NNY', showing that it is less likely to talk about the image after chatting (having two turns conversation) on off-image topics.

Overall, we could see that most of the utterances are still image-related. About 37% of utterances are not of the image-related topic, and 14% of the dialogue ('NNN') is on the topic totally not related to the given image. Looking into what constitutes the 14% non-image-related dialogue ('NNN'), we find that it is usually about stimulated from one of the objects in the image. For example, retrospecting the speaker's experience of committing suicide given an image of a building. This suggests that linking objects to much broader scenarios is an important direction for machines to reconstruct the utterance.

2.3.2 Image Information

Figure 2 shows the distribution of image information classes. 45.4% of the utterances contains the information from the given image. Among them, 13% of utterance has the exact match of image tags, and the 12% of image objects are mentioned but not in the form of the image tag. This result points that expanding the vocabulary size of the tags or

Score	Utterance	Style
0	If only they knew what awaits them outside, a world of happiness and bliss.	*Foolish*
1	Looks like daddy is ready to play his songs.	*Caring*
2	That's it, I going to Vegas tomorrow. Who's coming with me?	*Spontaneous*

Table 4: Examples of the degree of how much the utterance reflecting the given style.

adding synonyms of the tags might increase the accuracy of the reconstructed utterances. In the 12% utterances when the image object is referred by a pronoun, the utterances are usually the description of the image object, which means being able to capture the attributes of the objects in the image is essential for the utterance reconstruction. There is no correct tag name from the image object tags 8.3% of the time, which indicates that improving the object detection results might improve the utterance reconstruction up to 8.3%.

Besides the annotation results on the sampled data, we also calculate the exact match of the full validation set. As the result, 42.5% of utterances in the whole validation set contains the image tags (which is closed to our sampled result), and each utterance has in average 0.527 tags.

On the other hand, 23.2% of utterances are in O classes (O, OS, OP). These utterances have other image-related information that is not expressed in the object tags. This kind of information is usually the description of the event or scenario of the image. Thus, how to know the scene beyond the given objects is also important.

Moreover, there are 31% of utterances in class N do not contain any image information. These utterances are usually on the off-image theme and the only hint to reconstruct such utterances is from their conversational context, and there is nothing to do with the image information side.

To find out which type of image information is needed when the utterance theme is related to images, we calculate the ratio of classes conditioned on the image relevancy. The result is shown in the Image-related column of Table 5. We can see that the main difference is that the number of N class (utterance without image information) is largely decreased (-16.8%), and there are an distinctly increasing number of O class (other image info, +5.9%) and T class (image tag, +4.7%). This result further suggests that, to generate image-related utterances, the image information from image tags and related information is necessary.

Class	All		Image-related	
	Count	Ratio	Count	Ratio
N	34	31.48%	10	14.71%
O	19	17.59%	16	23.53%
T	14	12.96%	12	17.65%
TP	13	12.04%	10	14.71%
TN	13	12.04%	10	14.71%
TF	9	8.33%	6	8.82%
OS	5	4.63%	3	4.41%
OP	1	0.93%	1	1.47%

Table 5: Classification of the image information in the utterance. Image-related column is based on the image-related utterances.

2.3.3 Utterance Style

We score each utterance from 0 to 2 based on the relevance to the given style. The average score is 1.49, implying that the given style is the key factor to the generated utterance. Considering the style and using text generation models with style control module would be a favor for this task.

3 Augmenting Image Information

The analysis in previous section suggests the importance of the image objects and the non-object image-related information which is often the scenario of the image. Therefore, we augment the image feature by image caption to capture the scenario information and by texts of image objects so that the model could learn to copy and use the exact object tags. We also replace the single full-image feature in the baseline model with several image region features to facilitate the extraction of image object information.

3.1 Enhanced Image Features

3.1.1 Image Tags

As shown in (Li et al., 2020), attaching the object tags to the raw image feature improves the result on several language-image tasks such as image captioning. Current models integrate all information in the latent-level after text and image are encoded,

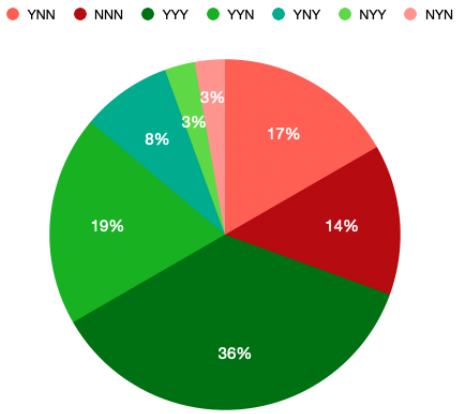

Figure 1: Different combination of image-related utterances in 3-turns dialogues. Y: image-related utterance, 63%; N: the utterance could exist independently of the image, 37 %. Green hue indicates the dialogue is more image-dependent (Y is more than N in a dialogue), the red family suggests the dialogue theme is more irrelevant to the image.

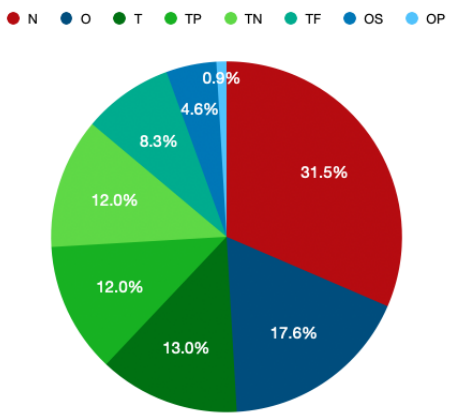

Figure 2: Classification of the image information in the utterance. Green hue refers to the existence of image objects information, blue hue refers to other image related information not in the image object tags, and red refers to the utterance without any image information.

but the decoder might tend to focus on the text encoder output and ignore the image encoder output. By explicitly listing out the objects in the image in text, even the decoder is not paying attention on the image encoder output, the model would still obtain the image information in addition to the dialogue.

3.1.2 Image Caption

Besides image objects, we also use a pretrained image captioning model to generate captions to represent the image in text. While there are some implicit knowledge that couldn't be captured by the objects, we use image captioning models to overcome this problem. For example, if an image contains a bride and a groom, the object detection model wouldn't output the keyword "wedding", but an image captioning model may be able to generate such keyword since it might often appear in the context. Another benefit of using image caption in the model is that the caption often offer additional image-relevant keywords. Despite grounding on a single image, a natural conversation with an image often goes beyond the image content and stretches out to relevant knowledge. The image caption could provide such related keywords for generating more engaging responses and continuing the conversation length.

3.1.3 Region Features

Furthermore, we use multiple region features in our model instead of a single feature for the whole image as in previous works (Shuster et al., 2020c; Ju et al., 2019). Our empirical result show that the model could better capture image relations by using multiple region features than a single image representation.

3.2 Experiments

We run our experiments on the ImageChat dataset (Shuster et al., 2020a) which is described in Sec 2.1. All our experiments are conducted using the ParlAI (Miller et al., 2017) framework.

3.2.1 Baseline

We mainly compare with the state-of-the-art multimodal dialogue system: Multimodal Blenderbot (MMB) (Shuster et al., 2020c), which is a 2.7B Transforer-based model with 2 encoder layers and 24 decoder layers. The embedding dimension is 2560, image feature dimension is 2048, and the FFN size of the Transformer decoder is 10240. MMB is pretrained on reddit dataset (Baumgartner et al., 2020) and COCO image captioning dataset (Chen et al., 2015), and finetuned on ImageChat dataset and multiple text-only diagloue datasets used in the BlenderBot (Roller et al., 2021) (BST+), including ConvAI2 (Dinan et al., 2019a), EmpatheticDialogues (Rashkin et al., 2019), Wizard of Wikipedia (Dinan et al., 2019b), and BlendedSkillTalk (Smith et al., 2020).

3.2.2 Implementation

We obtain image tags from Scene Graph Benchmark (Han et al., 2021) and the image caption from

pretrained VinVL model (Zhang et al., 2021). Instead of using a single image feature, we use multiple region features and encoded them to a single vector through multilayer perceptron. The image feature dimension is set to 2054, with additional 6-dim image information such as weight and height to the 2048-dim FasterRCNN feature in the original model. Each image is paired with 1 to 10 unique tags, an image caption with maximum 12 tokens, and at most 32 image object features.

We finetune the Reddit pretrained model on different datasets following the instruction from MMB[1] (Shuster et al., 2020c). Since the results in MMB shows that early fusion is slightly better, we also early-fuse image and text features. After the utterance, tag, and caption are embedded, the text embeddings are concatenated with the image feature and feed into a Transformer. Unlike in (Shuster et al., 2020c), where they use the coarse classification of styles (positive, neutral, negative), we use the original style name in our inputs.

3.2.3 Evaluation

Following previous works, we report the number of perplexity (PPL), Rouge-L, BLEU-4, and F1 score. As existing research has reported that these numbers are not highly correlated with human evaluation (Liu et al., 2016; Li et al., 2016), we also reported BERTScore (rescale) (Zhang* et al., 2020), which reflects the semantics similarity instead of the token-wised matching.

To show the engagingness of generated responses and the relevance to images, we run the image text retrieval task using VinVL (Zhang et al., 2021) pretrained model for image text retrieval. We also report the number of average length, unique vocabularies, and Distinct-1 (Li et al., 2015) (normalized unigram) of the utterances.

3.3 Results and Analysis

Table 6 demonstrates that our enhanced image features improves the strong baseline without training on additional image captioning and text-based dialogue datasets. Comparing the result between L1, L2 and L4, L5 in Table 6, we can see that removing the text-based dialogue datasets (BST+) degrades the result when there is no enhanced image features (denoted as *image*$^+$) provided (L1 to L2), but improves the performance when using *image*$^+$ (L3 to L4). This result implies that *image*$^+$ provides much more useful information than those text-only datasets (BST+) that neither additional text-only dialogue datasets nor image captioning pretraining is needed. Besides, a pipeline approach of explicitly adding image caption to the model is better than end-to-end training on the additional image captioning task.

We also found that the Reddit pretraining is essential for dialogue generation. Without pretraining, the perplexity would boost to about 34 and all other metrics get worse based on our empirical results. In fact, the perplexity is already around 26 at the very beginning of the training when finetuning on the Reddit pretrained model.

As shown in Table 7, we successfully demonstrate our model's superiority on generating responses that are more diverse and relevant to the images. With enhanced image features, we get the best retrieval result in both image-to-text and text-to-image retrieval and even outperform the human references, showing that our generated responses are more image-related. We also generate longer sentences with more diverse vocabularies than the MMB baseline.

We provide the dialogue outputs from MMB and our MMB + *image*$^+$ in the Appendix.

4 Related Work

To approach a more human-like chatbot which is capable of communicating with human naturally in open-domain topics, various research directions have been explored. Dinan et al. (2019a); Zhang et al. (2018) propose to assign chatbots with a consistent personality; Empathetic Dialogues(Rashkin et al., 2019) dataset was created to train chatbots to consider the feeling of people and display empathy; Dinan et al. (2019b) propose Wizard of Wikipedia, which aims at grounding the conversation on knowledge base. Roller et al. (2021) propose BlenderBot, a chatbot trained on a composition of the three aforementioned skills (personalization, empathy, and knowledge). These open-domain dialogue systems are text-based.

Image-Grounded Conversations (Mostafazadeh et al., 2017) and Visual Diagloue (Das et al., 2017) are the early proposed visual dialogue datasets. However, they and other image-grounded conversational datasets proposed afterwards (Kottur et al., 2019; Zhao and Tresp, 2018) are more similar to the multi-turn visual question answering task rather

[1] https://github.com/facebookresearch/ParlAI/blob/main/parlai/zoo/multimodal_blenderbot/README.md

L	Model	Datasets	PPL	Rouge	BLEU	F1	Bert Score		
							P	R	F1
0	MMB	R,I,C,B	12.64	18.00	0.418	13.14	-	-	-
1	MMB	R,I,C,B	13.60	12.40	0.386	12.94	33.81	25.21	29.49
2	MMB	R,I,C	15.00	11.35	0.278	11.81	31.73	23.52	27.61
3	MMB	R,I	12.89	13.04	0.419	13.52	32.58	24.23	28.39
4	MMB + *image*$^+$	R,I,C,B	**12.63**	**13.36**	0.447	13.75	34.76	26.36	30.54
5	MMB + *image*$^+$	R,I	12.76	13.29	**0.461**	**13.82**	**35.36**	**26.38**	**30.85**

Table 6: We compare the model pretrained on Reddit (R) dataset and finetuned on ImageChat (I), Coco Captioning (C), and BST+(B). *image*$^+$ refers to the enhanced image features (image tags, caption, and region feature). L0 is from the MMB paper (Shuster et al., 2020c), and L1 is the result of re-running evaluation script[2] on the model provided by MMB's authors.

Model	Image-to-Text		Text-to-Image		Length	Vocabs	Distinct-1
	R@1	R@10	R@1	R@10			
Gold	0.02	0.14	0.03	0.32	**9.90**	**9,431**	**0.064**
MMB.	**0.04**	0.16	0.03	0.29	7.87	3436	0.029
MMB + *image*$^+$	**0.04**	**0.26**	**0.04**	**0.35**	8.04	3865	0.032

Table 7: Result of image relevancy on validation set. Gold is the utterances by human.

than in the form of a natural conversation. Several video-grounded dialogue systems have been proposed (Alamri et al., 2019; Thomason et al., 2019; Le et al., 2021), but these works focus on task-oriented dialogue such as navigation, different from our goal of open-domain chitchat. ImageChat dataset (Shuster et al., 2020a) was released for research on conversation grounding on image and emotional style.

Ju et al. (2019) propose to combine the state-of-the-art image models and dialogue agent using the combiner that consists of multiple Transformers, while Shuster et al. (2020a) with similar architecture shows that their sum combiner suming over representations works better than their attention alternative. Shuster et al. (2020b) trained a multi-tasking model which consists of 12 subtasks to allow the agent to be capable of asking/answering question, having its own persona, and grounding on images and external knowledge. Shuster et al. (2020c) proposes Multi-Modal BlenderBot (MMB), which can be seen as BlenderBot with additional image features. In MMB, they study different image and text feature fusion methods instead of directly conducting multi-modal training, and examine the effect of fine-tuning strategies and domain-adaptive pretraining on image captioning dataset.

5 Conclusions

In this paper, we analyze the factors that influence the quality of open-domain conversation on images from three aspects: image relevancy to the conversation theme, image information in the conversation and style of utterance. We divide the conversations into image-related and non-related groups based on the conversational theme, and find that the image objects as well as the non-object related image information contribute to 56% and 29% of image-related dialogues; and the scenario that includes the image object is vital to infer the off-image topics for the non-image-related conversations. Also, around 30% image objects are missing in the tags, thus expanding the vocabulary sizes and improve the accuracy of the object detection model would help the response generation of conversations on images.

We propose to incorporate more image information, i.e., image tags and image captions, into the baseline model based on our analysis on the conversations on images. The result suggest that our enhanced image features help to generate more image-related and diverse conversation responses, verifying the effectiveness of our findings. We believe that our in-depth analysis and proposed findings would benefit the future research on the open-domain conversation on image.

References

D. Adiwardana, Minh-Thang Luong, David R. So, J. Hall, Noah Fiedel, R. Thoppilan, Z. Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

A. Das, Satwik Kottur, K. Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019a. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and J. Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv*, abs/1811.01241.

Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. 2021. Image scene graph generation (sgg) benchmark.

Da Ju, Kurt Shuster, Y-Lan Boureau, and J. Weston. 2019. All-in-one image-grounded conversational agents. *ArXiv*, abs/1912.12394.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 582–595, Minneapolis, Minnesota. Association for Computational Linguistics.

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 2021. Video-grounded dialogues with pretrained generation language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

N. Mostafazadeh, Chris Brockett, W. Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *IJCNLP*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020a. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 2414–2429, Online. Association for Computational Linguistics.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and J. Weston. 2020b. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. *ArXiv*, abs/1911.03768.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2020c. Multi-modal open-domain dialogue. *arXiv preprint arXiv:2010.01082*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Rui Zhao and Volker Tresp. 2018. Efficient dialog policy learning via positive memory retention. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 823–830. IEEE.

A Example Outputs

Irrational: Is she doing a backflip twist?
MMB: I don't think she's going to catch the ball.
Our: I bet they are going to fall and hurt themselves.

Obnoxious: I don't think so. I think she is passing gas.
MMB: She is doing a back flip.
Our: I hope she falls and breaks her neck.

Irrational: She needs to find a restroom quick!
MMB: She is definitely not passing gas.
Our: I think she is trying to fly.

Irrational: Is she doing a backflip twist?
MMB: I don't think she's going to catch the ball.
Our: I bet they are going to fall and hurt themselves.

Obnoxious: I don't think so. I think she is passing gas.
MMB: She is doing a back flip.
Our: I hope she falls and breaks her neck.

Irrational: She needs to find a restroom quick!
MMB: She is definitely not passing gas.
Our: I think she is trying to fly.

Curious: What did they have to do to help support canc er? Fundraisers always intrigue me
MMB: What is that kid doing?
Our: I wonder what they are doing.

Patriotic: Maybe they donated. All Americans should do nate to show that we are a proud and caring country
MMB: They are fighting for our country.
Our: They are supporting their country by raising mone y for cancer research.

Curious: They don't look american, I wonder if this is somewhere else?
MMB: I wonder what they did to help.
Our: I wonder what they did to help

Figure 3: Examples of the generated results from MMB(Shuster et al., 2020c) (underlined) and with our enhanced image features *image*$^+$ (bold).