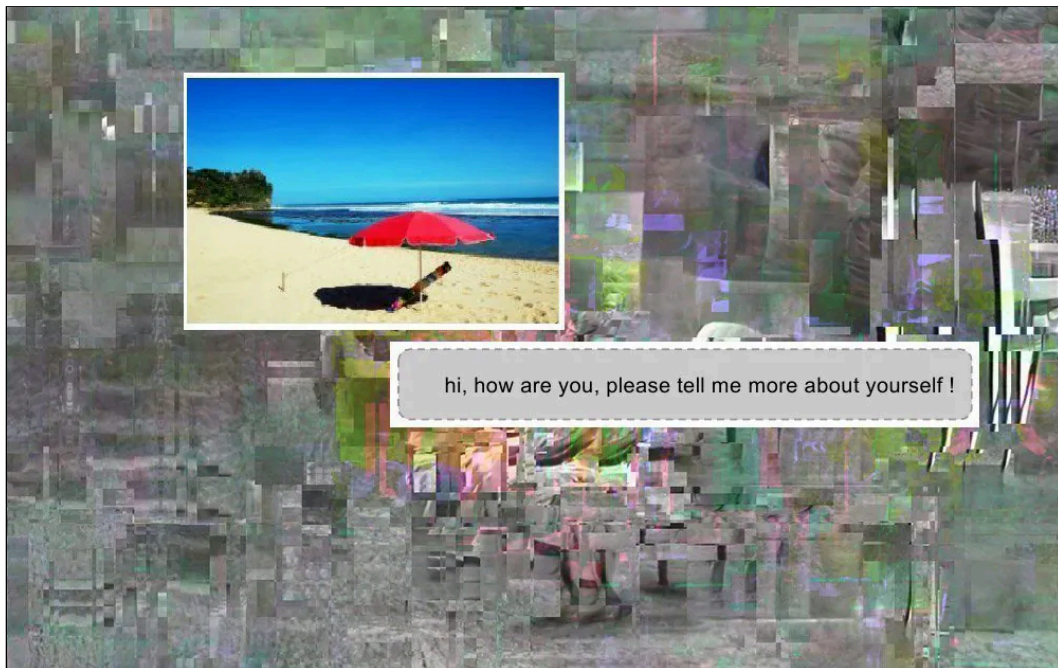


Teaching AI to Understand and Use Images in Dialogue

Originally published at <https://www.unite.ai/teaching-ai-to-understand-and-use-images-in-dialogue/>



Researchers from South Korea have developed a dataset designed to aid research into AI's understanding of the way that humans use images in dialogue, and to help natural language models to participate in this very recent development in human communications.

The [paper](#), from KAIST at Daedeok Innopolis, notes that research into such multi-modal dialogue systems over the last ten years has been hamstrung by datasets and methodologies centering on disciplines that are peripheral to the topic, such as [visual question answering](#) and [image captioning](#).

In these older approaches, images are evaluated out of the lexical context of a conversation, with no understanding of the way that the dialogue is enhanced and developed by image responses, and no cross-domain schema for decoding the contributions of visual contributions to discourse.

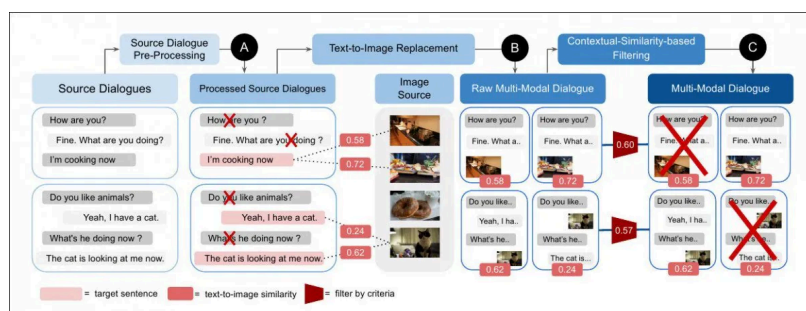
Images as First-Class Facets of Dialogue

Many of the aforementioned approaches to date have been initiatives or developments from Microsoft's AI research arm, which in 2017 also [examined](#) the topic of multimodal conversations that are *begun* by an image, rather than freely using images as dialogue components.

To address the shortfall in research data, the South Korean researchers have developed a dataset of 45,000 dialogue instances involving the ad hoc usage of images, without concentrating on [viral 'meme' images](#): the latter, though an area of interest in language research, is arguably less of a challenge, because the meaning of viral memes [can be inferred](#) more easily through thousands of in-context uses on social media platforms.

Developing Illustrations as a Substitute for Text

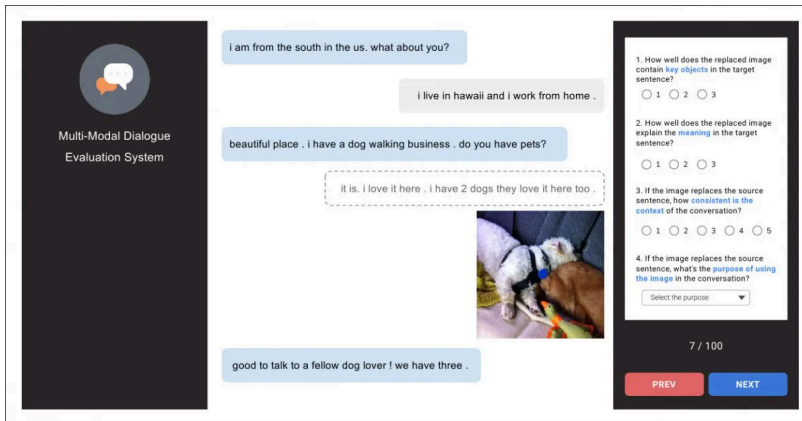
In order to develop a methodology for word/phrase>image bilateral transliteration, the South Korean researchers have trained a machine learning system to substitute parts of a text-based conversation into semantically relevant image content.



Architecture of the Korean system for generating a dataset for multimodal dialogue research. Source: <https://arxiv.org/pdf/2107.08685.pdf>

Pre-processing the target phrases involved the deletion of [stop words](#) that might inhibit prediction of the following sally in the conversation, and the pruning of inferior quality exchanges via contextual similarity filters.

To test the utility of the dataset, the researchers set a module to predict the next 'turn' in the dialogue while considering the context of the conversation and the images involved.



The human evaluation GUI used in the research.

Five external datasets were used as base material for the 45k dataset (which is [available on GitHub](#)). Three are text-based elements: [DailyDialog](#), a manually-annotated multi-turn text-based set from 2017; and Facebook's [EmpatheticDialogues](#) and [PersonaChat](#), both from 2018. The two image-based datasets used were MS-COCO and [Flicker30k](#).

```
morning how old are you i am 34", "i am much older than that . my grandchildren are in their thirties", "do you have
two dogs are my everything .", "no pets . i am retired so i love to travel . pets would slow me down", "i understand
idea of travel is a long hot bubble bath .", "i used to take a lot of bubble baths when i was recovering from cancer"
cancer get cured ?", "i did ! feels great to be alive", "that is wonderful ! do you play chess ?", "i do not . do you
it is one of my hobbies . what do you do for fun ?", "i spoil my grandkids . they want me to get a smartphone , do yo
should ?", "i think it depends on the kid . however all or none . are you married ?", "i am . are you married ?", "no
i . some day the right person will come along .", "do you date much ? you have to put yourself out there", "replaced i
'img_idx': 62, 'score': 0.556051395101951, 'dialog_dataset': 'persona', 'dialog_file': 'chat10043.txt', 'img_dataset'
'img_file': 'COCO_val2014_000000283210.jpg', {'dialog': ['hello , how are you this evening ?', 'hello how are you to
doing well , thank you . playing online chess , how are you ?', 'i am chillin , do you have any pets ?', 'i have two
are watching my bath run .", "i like chocolate , do they ?", "oh no , chocolate is bad for dogs .", "why i am only 4
akes them very sick .", "change of subject , what is your favorite color ?", "i love purple , what is yours ?", "i l
reminds me of summer time .", "summer is my favorite season .", "i have a best friend", "i do too , i have known mine
hirty years .", "i really like waffles too", 'replaced_idx': 4, 'score': 0.553968126620195, 'dialog'
'persona', 'dialog_file': 'chat10044.txt', 'img_dataset': 'coco', 'img_file': 'COCO_val2014_000000283210.jpg'], ('dia
, how are you this morning ?', "not good . things in my life are not going well .", "i am so sorry to hear that . hav
listening to taylor swift ?", "music is not going to bring my wife and kids back !", "it will help with your depressi
music does this for me .", "since they left all i do is work and play games on facebook .", "do you work ? i am a man
convenience store .", "i do have a boring job . do you go to church ? i am a non believer .", "no , i am pagan . natu
church .", "maybe my wife will come back if i lose weight . any suggestions ?", "cardio and weight lifting . as for y
ounseling may help .", "i used to ride bikes with my kids alot . do you have kids ?", "no , i do not . only dogs .",
heard pets can help with depression . what breeds do you own ?", "springer spaniel and a labrador retriever .", "do t
happy ?!", 'replaced_idx': 12, 'img_idx': 62, 'score': 0.5753145261984894, 'dialog_dataset': 'persona', 'dialog_file'
```

Image/text pairs – JSON schema of phrases in the dataset, associated with images (in this example) from Microsoft's COCO image database.

Text to image replacement for the system was powered by the pre-trained [Visual Semantic Reasoning Network](#) (VSRN), developed in 2019 out of Northeastern University at Boston. VSRN was set to operate on manually pre-selected phrases from the contributing text datasets.

Establishing Coherence

Coherence of the source datasets was established by developing six combinations of each dialogue dataset, correlated to instances in each image dataset, and evaluated over several rounds by humans.

The human scoring was based on three criteria: consistency to the context of the exchange; image-relevance to the core concept the image was trying to express; and the extent to which the image contained key objects from the target sentence.

Considering the latter criteria, it could be argued that the schema the researchers decided on has largely discounted the possibility of humorous, sarcastic, abstract or metaphysical possibilities for the semantic meaning of an image that might be injected into a text conversation.

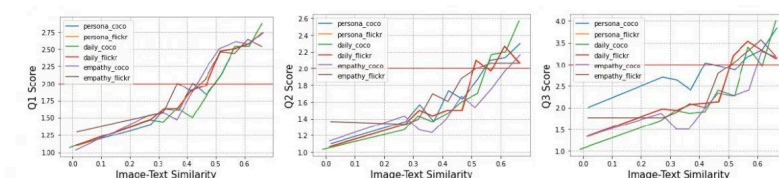
However, this is seminal work, and it has to start somewhere, while considerable efforts are being expended elsewhere in the Natural Language Processing (NLP) sector to [map instances of sarcasm](#), among other less tangible examples of the image/text relationship.

Testing

To test the data generation framework, the researchers used a three-part retrieval model based on Facebook's 2020 [Image-Chat](#) research. The module comprises [Resnext-101](#) as an image encoder; Google's [BERT](#) for the text encoder; and a custom fusion module for these.

The system achieved 50.35 and 14.38 on the current and next sentence prediction task, improving on the baseline for each task.

Later, two researchers were tasked with creating 100 multimodal dialogues by inserting images into conversations manually, and running the system against these 'organic' multimodal conversations. The system was able to predict current and next-turn exchanges with high awareness of context even for these ad hoc examples.



Results of the testing for the Korean multimodal dataset generation system, revealing consistently high correlation between text-to-image similarity and human-based data.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More