Maria: A Visual Experience Powered Conversational Agent

Zujie Liang[1*], Huang Hu[2*], Can Xu[2], Chongyang Tao[2], Xiubo Geng[2], Yining Chen[2], Fan Liang[1] and Daxin Jiang[2]

1. Sun Yat-sen University, Guangzhou, China
2. Microsoft STCA NLP Group, Beijing, China

Code&data link: https://github.com/jokieleung/Maria

Association for
Computational
Linguistics
ACL-IJCNLP 2021

Motivation

- Most of existing chatbots are only trained on textual corpora, and lack of visual perception to the physical world
- Human conversations involve the visual association
- Co-occurrence relationship of the fine-grained objects on images reflects a kind of knowledge that can be hardly captured in traditional knowledge bases

Human-A: Hey! How was your vacation?

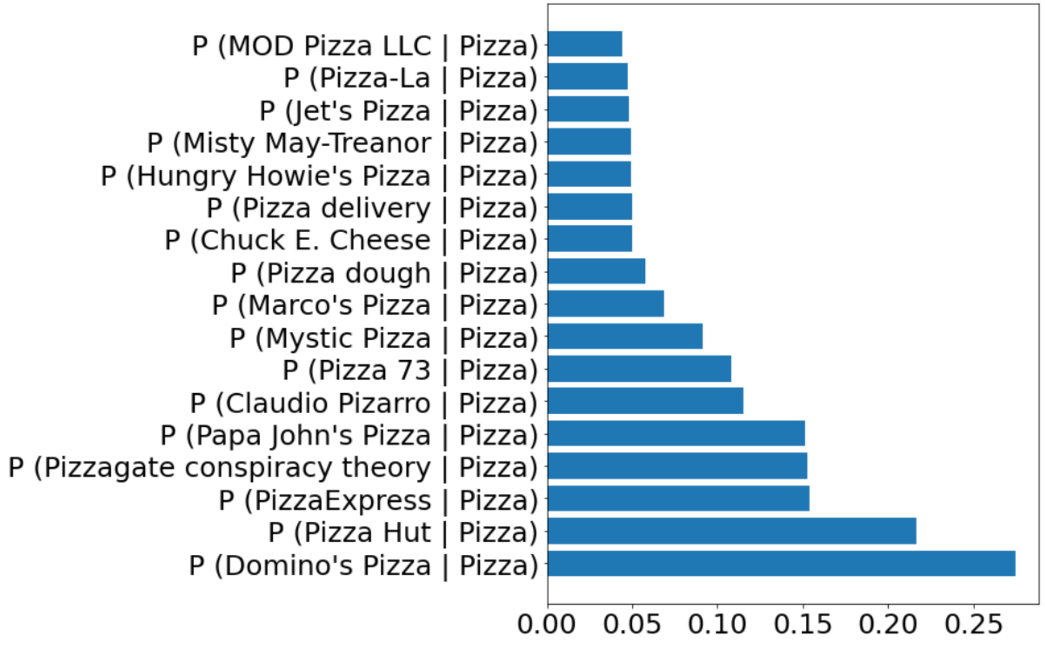
Human-B: *Awesome! I had a good time with my friends in Hawaii, the beaches are very beautiful there.*

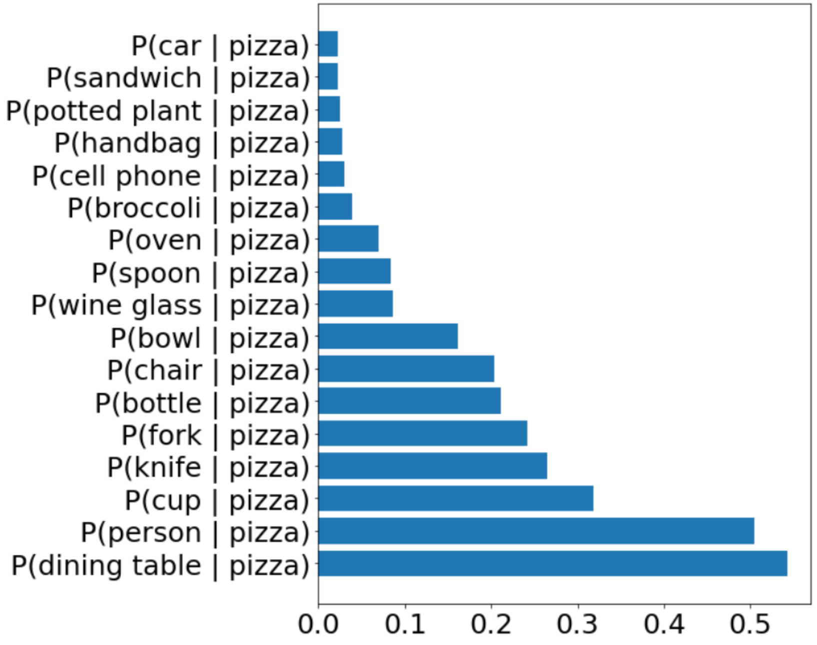
Human-A: Cool! did you play *beach volleyball* with your friends?

(**Human-A:** Cool, have you had a *BBQ* with your friends on the beach? The *grilled fish* was great!)

Human-B: *Nope, but it sounds great. Maybe next time.*

Example – “Pizza”


Item Co-occurence Distribution on Knowledge Graph


Object Tag Co-occurence Distribution on Images

Challenge

- Existing works on image grounded conversation (IGC) are constrained by the assumption that the crowd-sourced dialog is conducted center around a given image

- Existing methods on IGC are lack of the fine-grained understanding for image data

- How to effectively inject the visual knowledge extracted from image data into dialog model, and enable it to generate more informative and vivid responses

Contribution

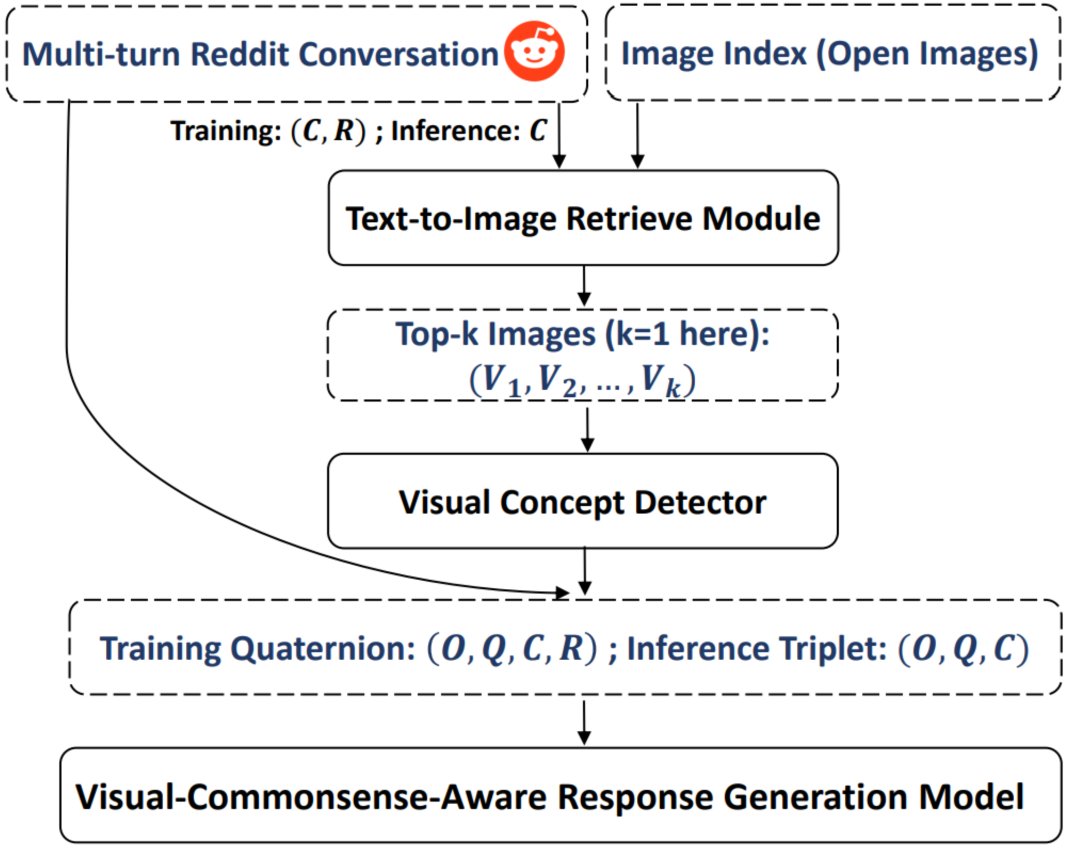
- To the best of our knowledge, this work is the first attempt to introduce the visual commonsense extracted from image data into open-domain dialog system

- Present Maria, a neural conversational agent consisting of three components, i.e., text-to-image retriever, visual concept detector and visual-knowledge-grounded response generator

- Propose a unified neural architecture for multimodal understanding and unimodal response generation

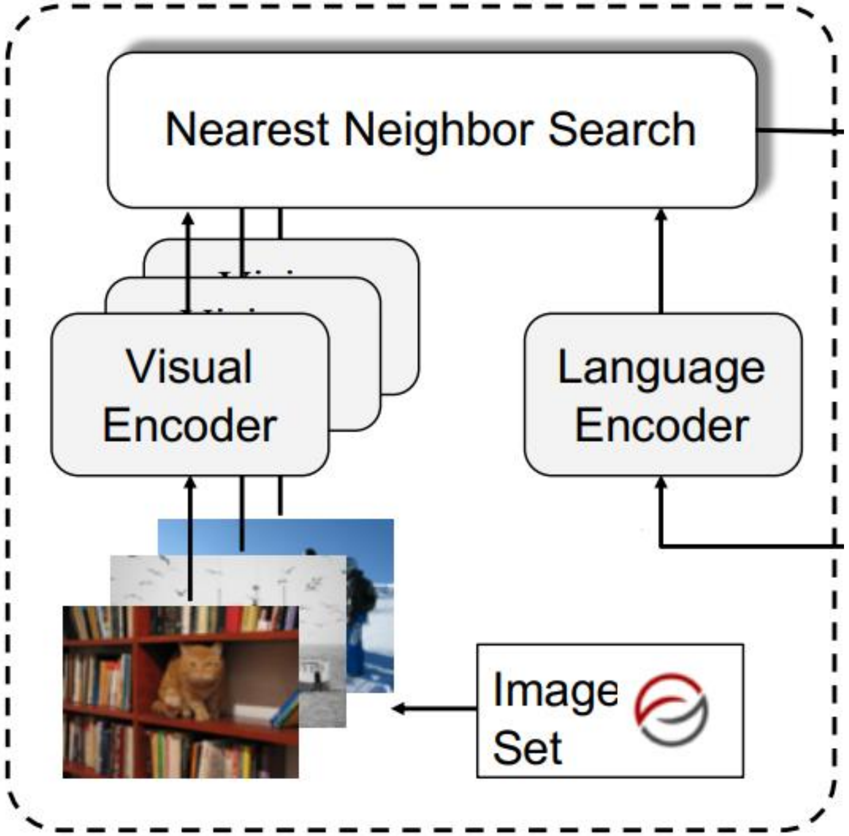
Framework

- Multi-turn Reddit Conversation Corpus [Dziri et al., 2019]
- Image Index
  - Open Images dataset [Kuznetsova et al., 2018]
- Choose top-1 image


Multi-turn Reddit Conversation
Image Index (Open Images)
Training: (C, R) ; Inference: C
Text-to-Image Retrieve Module
Top-k Images (k=1 here):
$(V_1, V_2, \ldots, V_k)$
Visual Concept Detector
Training Quaternion: (O, Q, C, R) ; Inference Triplet: (O, Q, C)
Visual-Commonsense-Aware Response Generation Model

Text-to-Image Retriever

- Model
  - Reproduction of text-to-image retrieval model in Vokenization [Tan et al., 2020] which learns a cross-modal retrieval model from sentence-image aligned data.
  - Two-stream architecture (faster than the single-stream) trained on **MS-COCO training set (113K)**: ResNext101 for visual encoder and BERT for language encoder

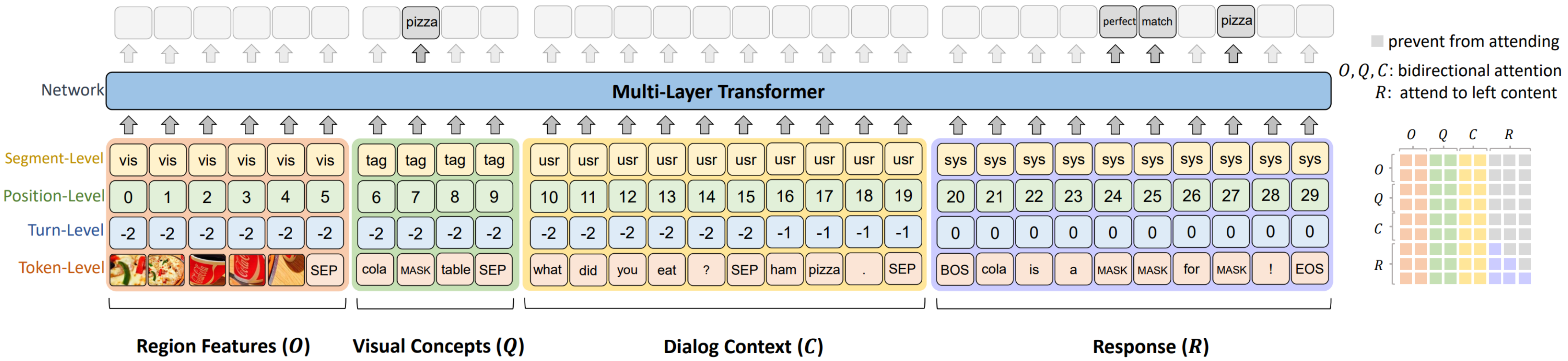


Visual Concept Detector

- Existing IGC works utilize the naive approaches, i.e., CNN-based models to extract the latent image features

- Introduce the pre-trained object detector to extract the **fine-grained image regions** and **corresponding concept labels**, and thus help Maria to better understand image details:
  - Model architecture: Faster-RCNN [Ren et al., 2015]
  - Trained on Visual Genome dataset [Krishna et al., 2017]

Visual-Knowledge-Grounded Response Generator

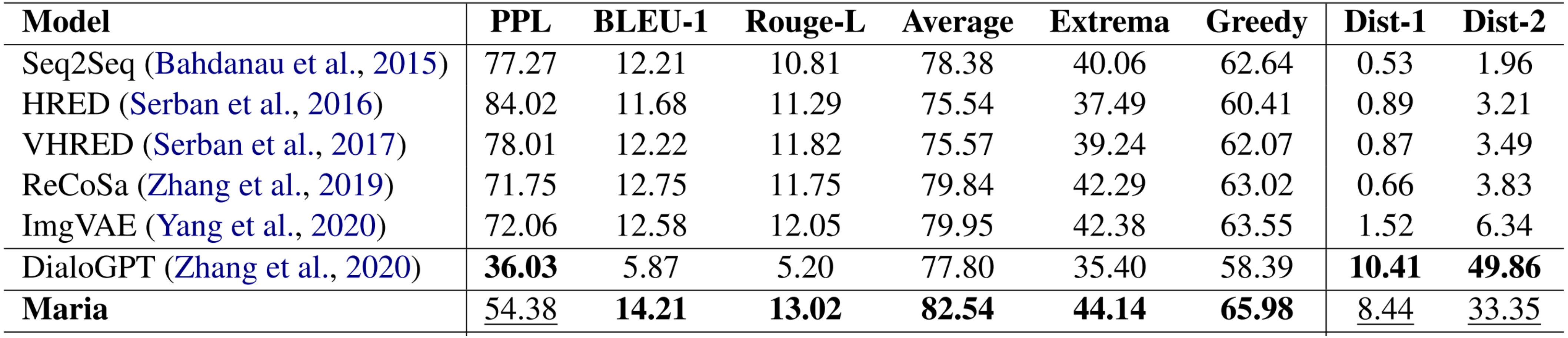


- Input Representation: token-level, turn-level, position-level, segment-level
- Mask Concept Prediction (MCP): multimodal understanding task, 15% tags
- Mask Response Prediction (MRP): unimodal response generation, 70% tokens
- MCP -> bidirectional self-attention, MRP -> attend all tokens in (O,Q,C) and leftward tokens in R

Dataset

- Reddit Conversation Corpus [Dziri et al., 2019]
  - Each dialog has 3~5 utterances, and the training/validation/test set has 1M/20K/20K dialogs
- Image Index
  - Sample 500K images from Open Images dataset [Kuznetsova et al., 2018]
- Utilize retrieval model to assign each dialog with a most relevant image, and extract visual region features and object tags by visual concept detector. Finally, construct *(bbox, tag, context, response)* 4-tuple training data

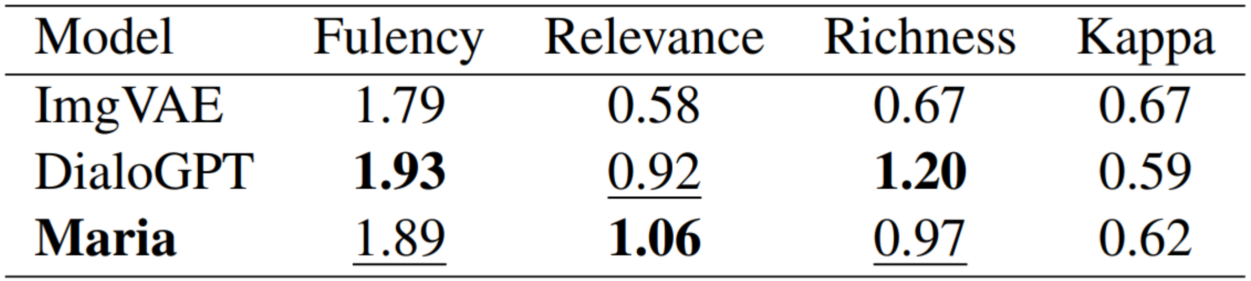
Automatic Metrics



Model	PPL	BLEU-1	Rouge-L	Average	Extrema	Greedy	Dist-1	Dist-2
Seq2Seq (Bahdanau et al., 2015)	77.27	12.21	10.81	78.38	40.06	62.64	0.53	1.96
HRED (Serban et al., 2016)	84.02	11.68	11.29	75.54	37.49	60.41	0.89	3.21
VHRED (Serban et al., 2017)	78.01	12.22	11.82	75.57	39.24	62.07	0.87	3.49
ReCoSa (Zhang et al., 2019)	71.75	12.75	11.75	79.84	42.29	63.02	0.66	3.83
ImgVAE (Yang et al., 2020)	72.06	12.58	12.05	79.95	42.38	63.55	1.52	6.34
DialoGPT (Zhang et al., 2020)	**36.03**	5.87	5.20	77.80	35.40	58.39	**10.41**	**49.86**
Maria	54.38	**14.21**	**13.02**	**82.54**	**44.14**	**65.98**	8.44	33.35

- **PPL:** DialoGPT finetunes GPT-2 on massive Reddit data (147M dialogs) while Maria is just trained on only 1M dialogs
- **Dist-1/2:** DialoGPT introduces an additional reverse model P(Context | Hypothesis) to rerank generated responses, thus improves the diversity of responses.

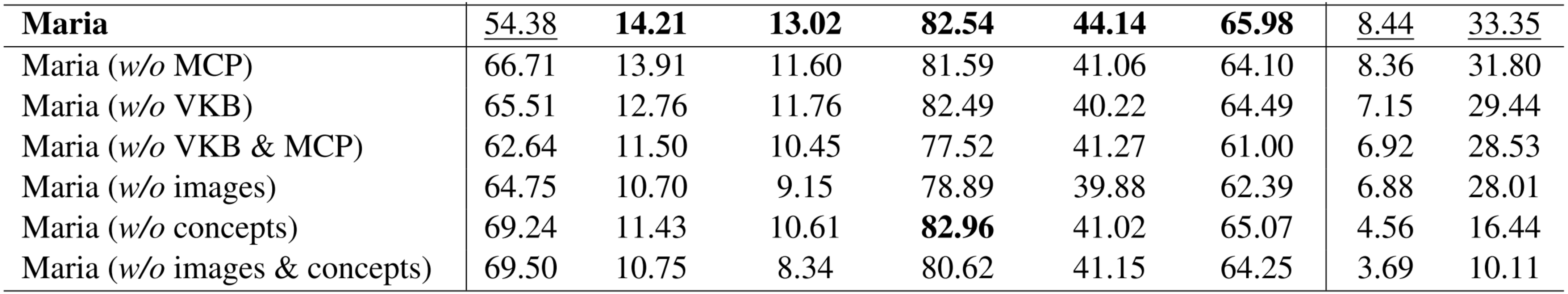
Human Judgement



Model	Fulency	Relevance	Richness	Kappa
ImgVAE	1.79	0.58	0.67	0.67
DialoGPT	**1.93**	0.92	**1.20**	0.59
Maria	1.89	**1.06**	0.97	0.62

- Three human annotators to score the response quality {0,1,2} on randomly 100 generated samples with respect to **Fluency**, **Relevance** and **Richness**
- The discrepancy of data distributions between training data (i.e., Image-Chat) and test data (i.e., Reddit Dialogs) of text-to-image synthesis model in ImgVAE limits its performance
- Maria introduce the extra image information by retrieval, which is the possible reason why it slightly outperforms DialoGPT on **Relevance**.

Ablation Study



Maria	54.38	**14.21**	**13.02**	**82.54**	**44.14**	**65.98**	8.44	33.35
Maria (*w/o* MCP)	66.71	13.91	11.60	81.59	41.06	64.10	8.36	31.80
Maria (*w/o* VKB)	65.51	12.76	11.76	82.49	40.22	64.49	7.15	29.44
Maria (*w/o* VKB & MCP)	62.64	11.50	10.45	77.52	41.27	61.00	6.92	28.53
Maria (*w/o* images)	64.75	10.70	9.15	78.89	39.88	62.39	6.88	28.01
Maria (*w/o* concepts)	69.24	11.43	10.61	**82.96**	41.02	65.07	4.56	16.44
Maria (*w/o* images & concepts)	69.50	10.75	8.34	80.62	41.15	64.25	3.69	10.11

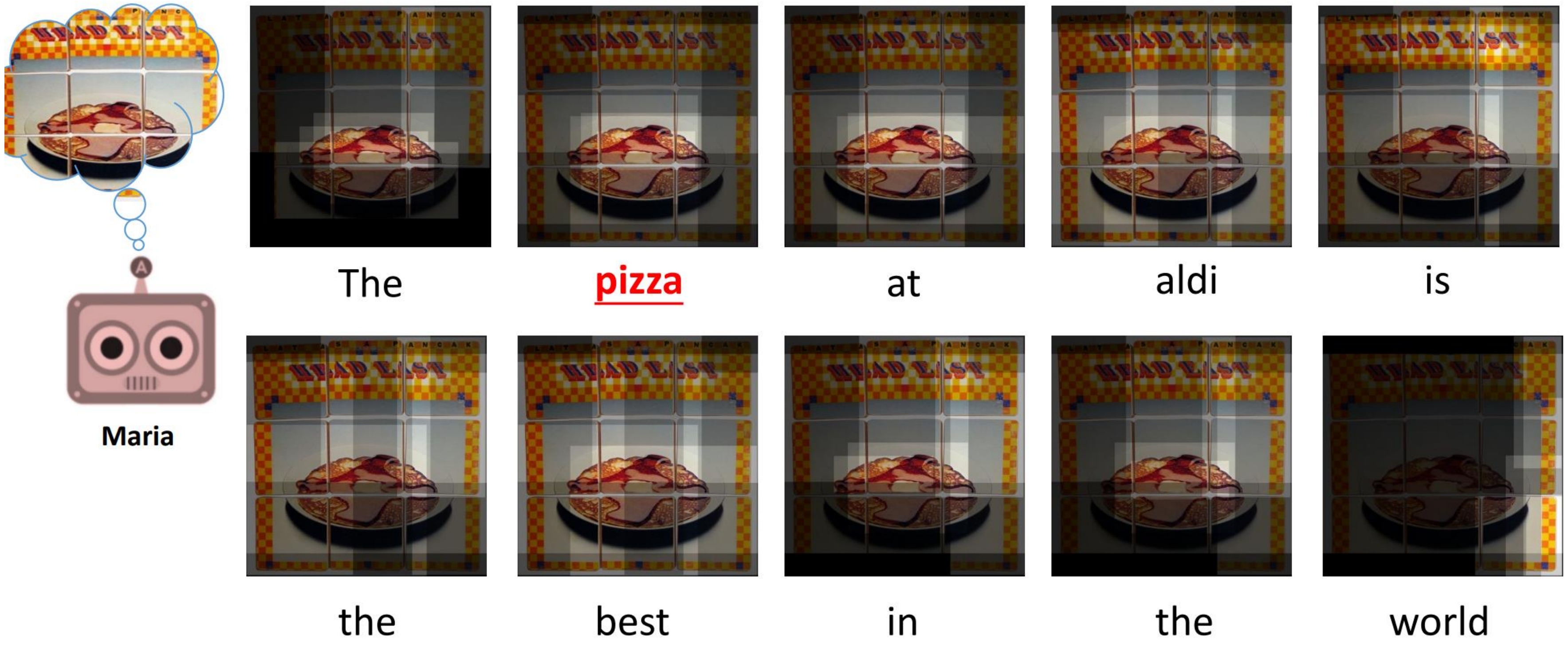
- Removing the extracted image regions & visual concepts leads to the dramatic performance drop on **diversity** (Dist-1/2)
- VKB & MCP can further improve the **relevance** and **diversity** of generated responses

Visualization Case

Dialog Context: **A:** No **Aldi**? hahah jokes.

B: **Aldi** is by far the best.

(Note: **Aldi** is the name of a supermarket)



Case Study

Dialog Context:

A:the pass of the **world cup** (james rodriguez)

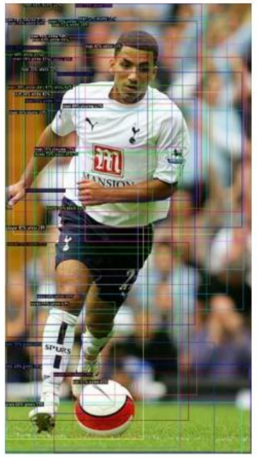
B: absoleutely gorgeous.

Baseline Response:

ImgVAE: I love the world cup.

DialoGPT: it 's not a pass of the world cup.

Maria



Detected Concepts: ball, socks, field, shirt, grass, man, shorts, ***player***, sock

Response: he's the best ***player*** in the world

Case Study

Dialog Context:

A: you 've just ruined my canada weather fetish , that it would be ideal place for me to live: freezing cold winters in mild summers.

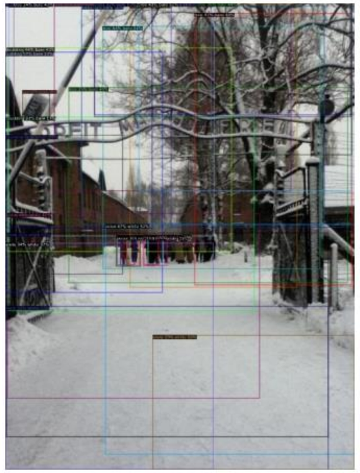
B: ya , there are parts of canada with very nice mild summers , but those same places have mild winters too.

Baseline Response:

ImgVAE: i 'm from canada , but i 've never seen a cold winters winters in cold winters winters .

DialoGPT: i 'm from the u.s. and i 've never heard of this.

Maria



Detected Concepts: person, fence, ground, tree, building, ***snow***, border, pole, gate

Response: i'm from canada and i've never been to a place that doesn't have ***snow***

Thanks

Code&data link: https://github.com/jokieleung/Maria