Data
Science
Innovation
Lab.

Supervisors:
Dr. Mohammad Akbari
Dr. Ali Mohades

Multimodal Task-oriented Dialog System

Presenter:
Amir Hossein Karimi

AGENDA

01 Introduction

02 Dataset

03 Related Works

04 Future Works

01
Introduction

introduction of the multi-modal dialogue system

Motivation

Photo by Peter Nguyen on Unsplash

Travel

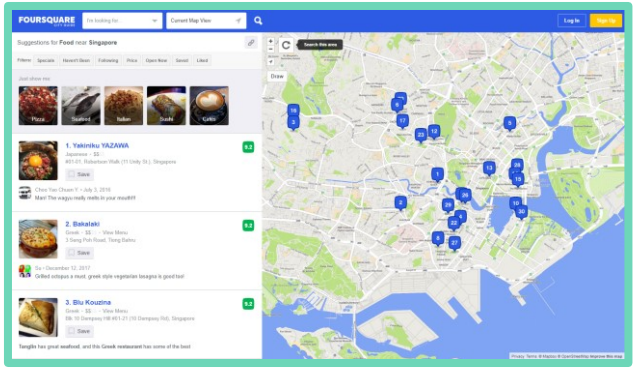
Suppose you travel to a new place.

How do you choose a restaurant?

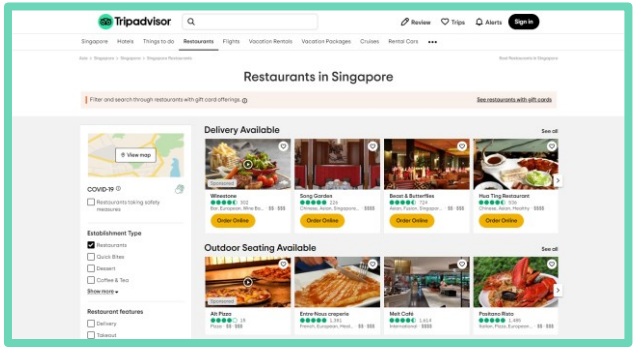
Which hotel do you book?

How do you explore the new place?

Motivation



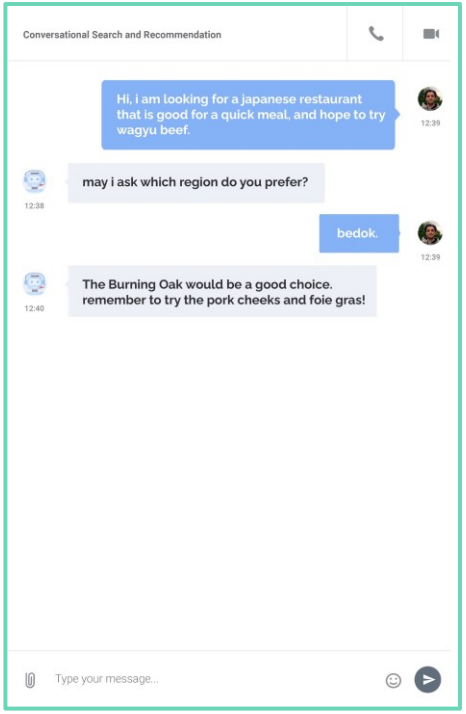
Screenshot from Foursqure.com and Tripadvisor.com



Searching the internet is one option.

But its very time consuming.

Motivation



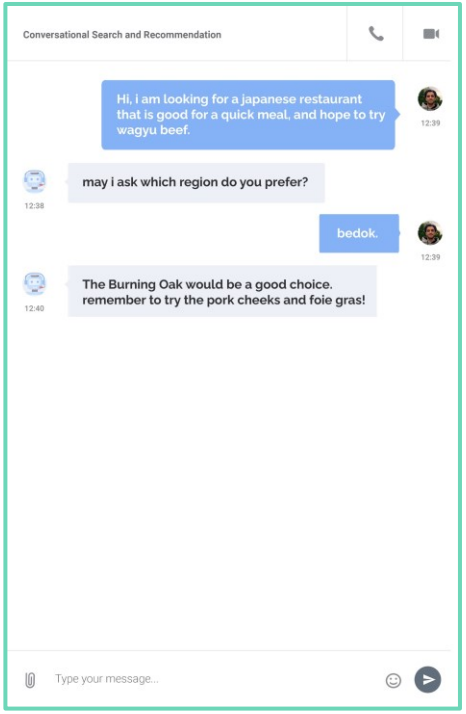
Picture by the author

Another possibility is to use recommender systems.

These systems can determine the **optimal choice** by asking a few questions of the user.

However, the majority of current dialog-based systems are single-modal and **text-based**.

Motivation

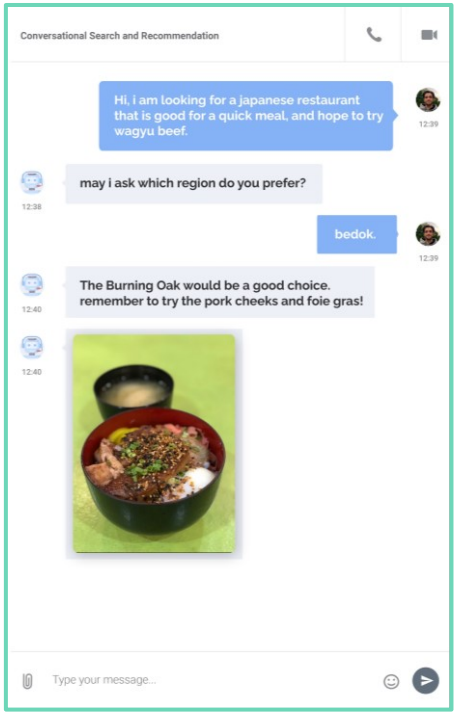


Picture by the author

How does the restaurant or the food have seemed?

Users have no idea what they may encounter.

Motivation



Picture by the author

A **multimodal dialogue system** can address this issue by displaying an image while making a recommendation.

MMTOD Tasks

Dialog State Tracking

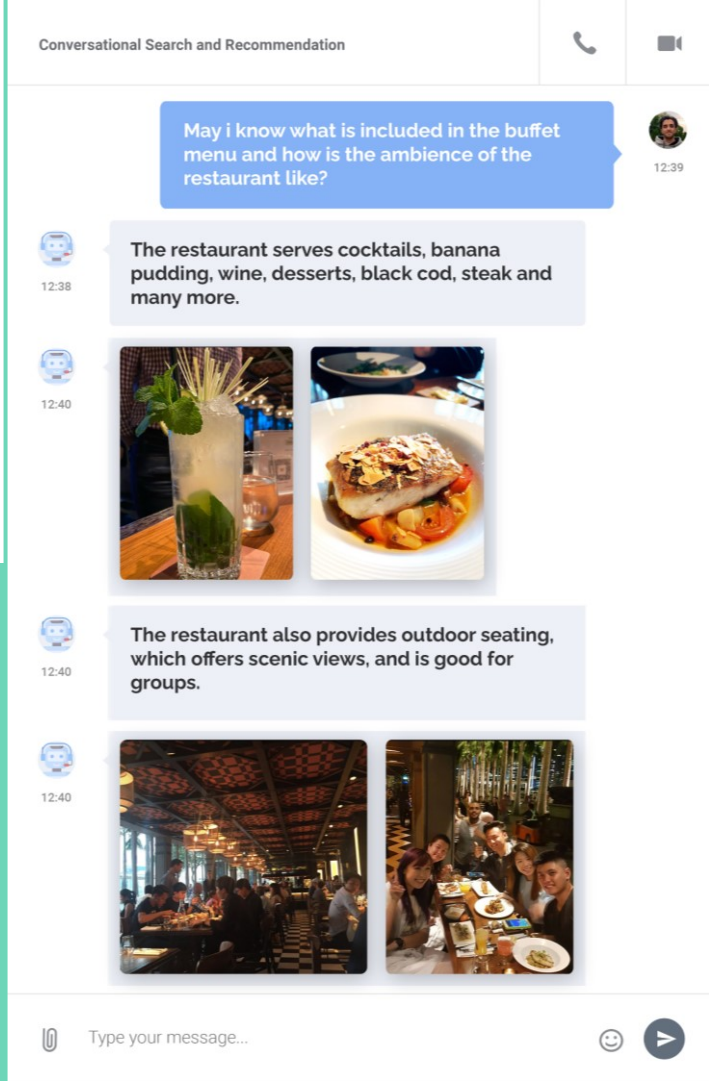
Monitor conversation actions and slot pairings

Action prediction

Correctly determine actions (e.g. Identify the right time to recommend)

Response generation

Generate fluent responses showing related photos



02
Dataset and Challanges

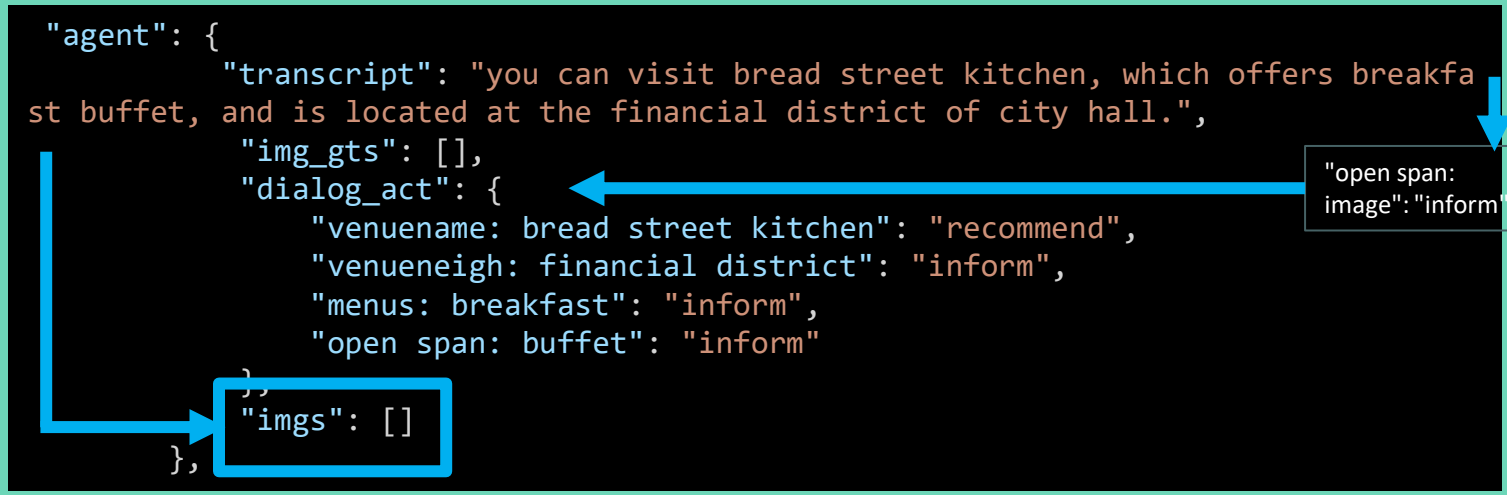
MMConv and Multimedia Challenges

Dataset

Datasets	# Dialogues	# Utters	Types	Domains	User Data	Modality	State Label
Facebook Rec [8]	1M	6M	Conv. Rec.	Movie	×	Text	×
REDIAL [17]	10K	163K	Conv. Rec.	Movie	×	Text	×
TG-ReDial [44]	10K	129K	Conv. Rec.	Movie	√	Text	×
OpenDialKG [23]	15K	143K	Conv. Rec.	Movie, book	×	Text	×
DuRecDial [21]	10K	156K	Conv. Rec.	Movie, music, news etc.	√	Text	×
MGConvRex [40]	7K	73K	Conv. Rec.	Restaurant	√	Text	√
WOZ 2.0[25]	1.2K	12K	Conv. Search	Restaurant	×	Text	√
DSTC2 [38]	1.6K	23K	Conv. Search	Restaurant	×	Text	√
FRAMES [9]	1.3K	20K	Conv. Search	Flight, hotel, budget	×	Text	√
KVRET [10]	3K	15K	Conv. Search	In-car assistant	×	Text	×
MultiWOZ [3]	8K	115K	Conv. Search	Hotel, restaurant etc.	×	Text	√
VisDial [5]	123K	2.4M	Image-based QAs	Concepts in image	×	Multi.	×
GuessWhat [6]	155K	1.6M	Image-based QAs	Concepts in image	×	Multi.	×
IGC [24]	4K	25K	Image-based QAs	Concepts in image	×	Multi.	×
MMD [29]	150K	6M	Fashion Search	Fashion	×	Multi.	×
MMConv	5.1K	39.7K	Conv. Search	5 domains in travel	√	Multi.	√

Liao et al. “MMConv: An Environment for Multimodal Conversational Search across Multiple Domains”. (2021)

Dataset



Liao et al. “MMConv: An Environment for Multimodal Conversational Search across Multiple Domains”. (2021)

Challenge - Images

Images in datasets may be used for a variety of purposes, including:

- Take a look at the meal and its contents.
- Find places and foods that are similar to the user's image.
- Describe the space and mood of different locations.

These annotated datasets, such as DSTC2 and MultiWOZ, are **all text-based** saerehw , .snoitatonna fo gnikcal lla era ,DMM dna ,CGI ,laiDsiV sa hcus ,stesatad ladomitlum

Challenge - Images

Text:

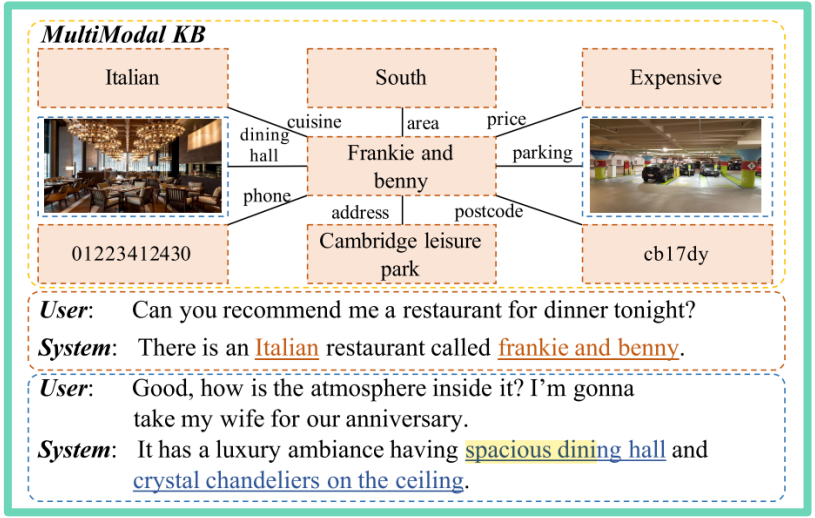
“Awesome lighting play on Christmas and new year! It could be **romantic** but it is too crowded, plus humid though. Nice to just sit together with your loved one and dream on your future plans together.”

Image:

Gardens by the Bay, retrieved from Foursqure.com

03
Related works

UniMF

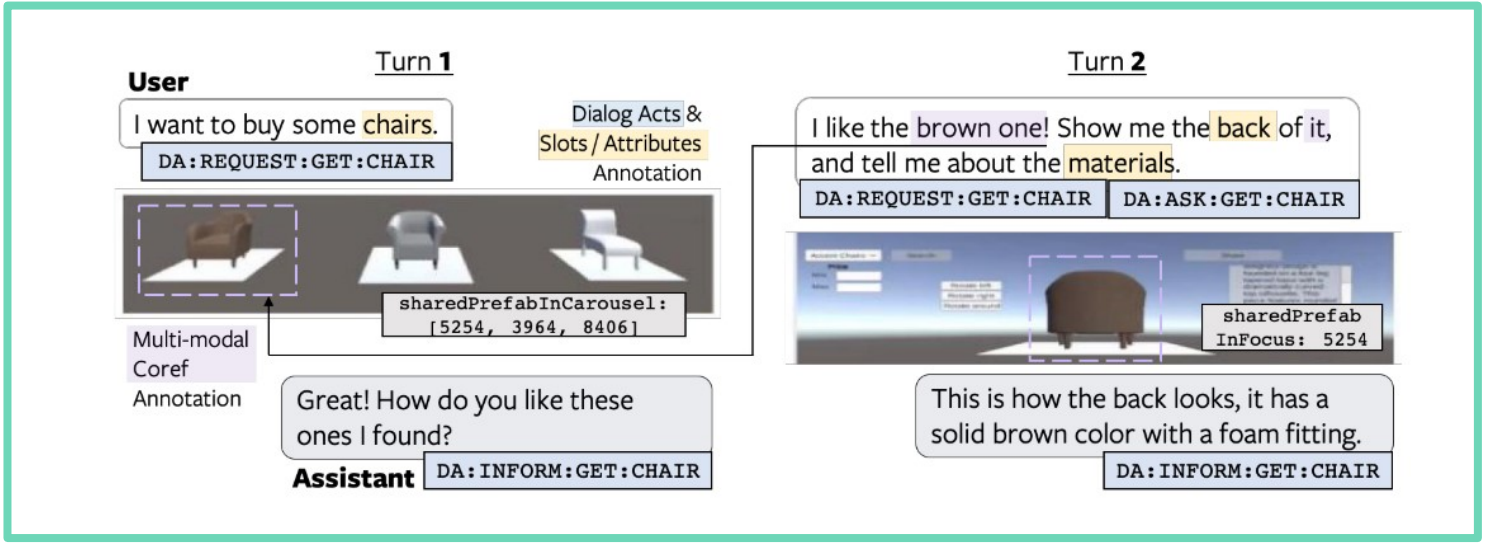


Yang, et al. "UniMF: A Unified Framework to Incorporate Multimodal Knowledge Bases into End-to-End Task-Oriented Dialogue Systems". IJCAI 2021 .

MMDialKB dataset:

Used MultiWOZ and search for place names in Google and add its image to the dataset

SIMMC



Moon, S., Kottur, S., Crook, P., De, A., Poddar, S., Levin, T., ... Geramifard, A. (2021). Situated and Interactive Multimodal Conversations.

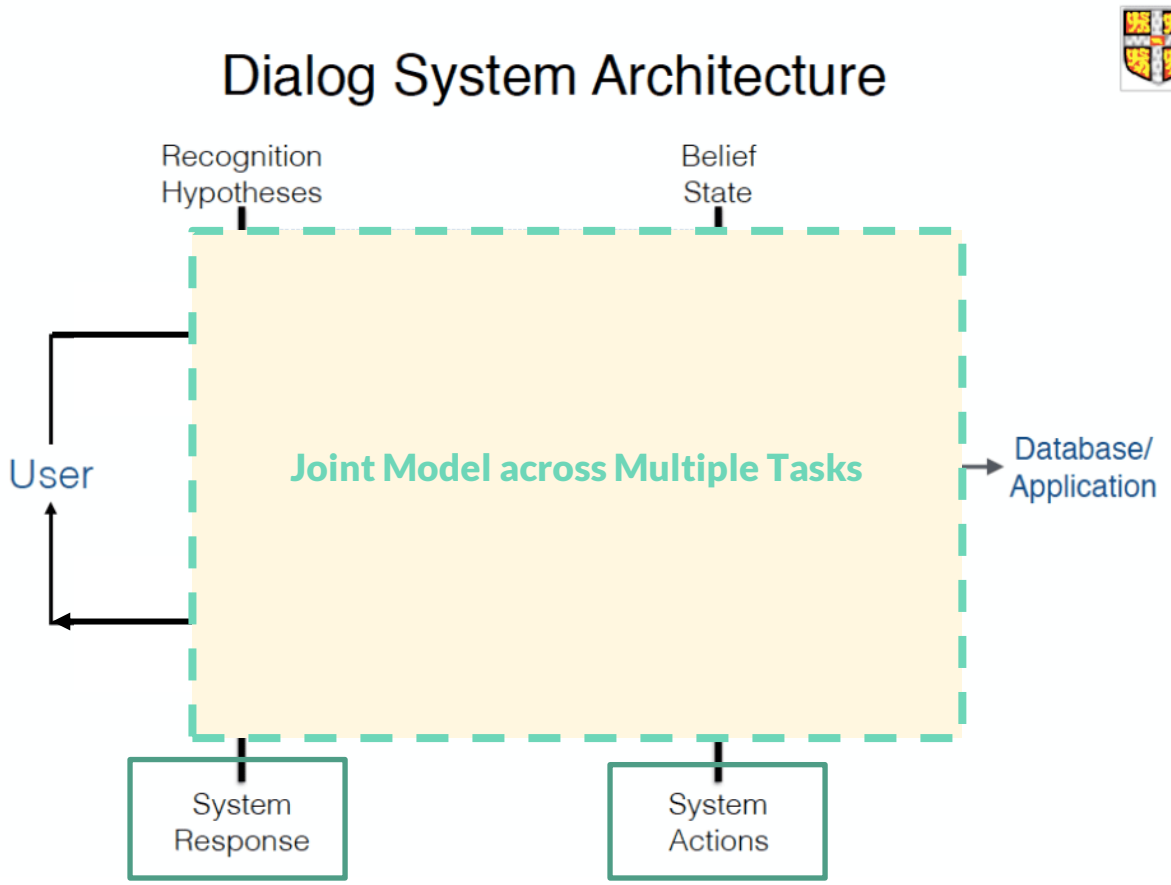
Other Related Works

- Le, Hung, S. Hoi, Doyen Sahoo, and N. Chen. "End-to-end multimodal dialog systems with hierarchical multimodal attention on video features." In *DSTC7 at AAAI2019 workshop*. 2019.

- Li, Zekang, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. "Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 2476-2483.

- Cui, Chen, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. "User attention-guided multimodal dialog systems." In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 445-454. 2019.

Other Related Works



SimpleTOD

Hosseini-Asl, E., McCann, B., Wu, C. S., Yavuz, S., & Socher, R. (2020). A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, *33*, 20179-20191.

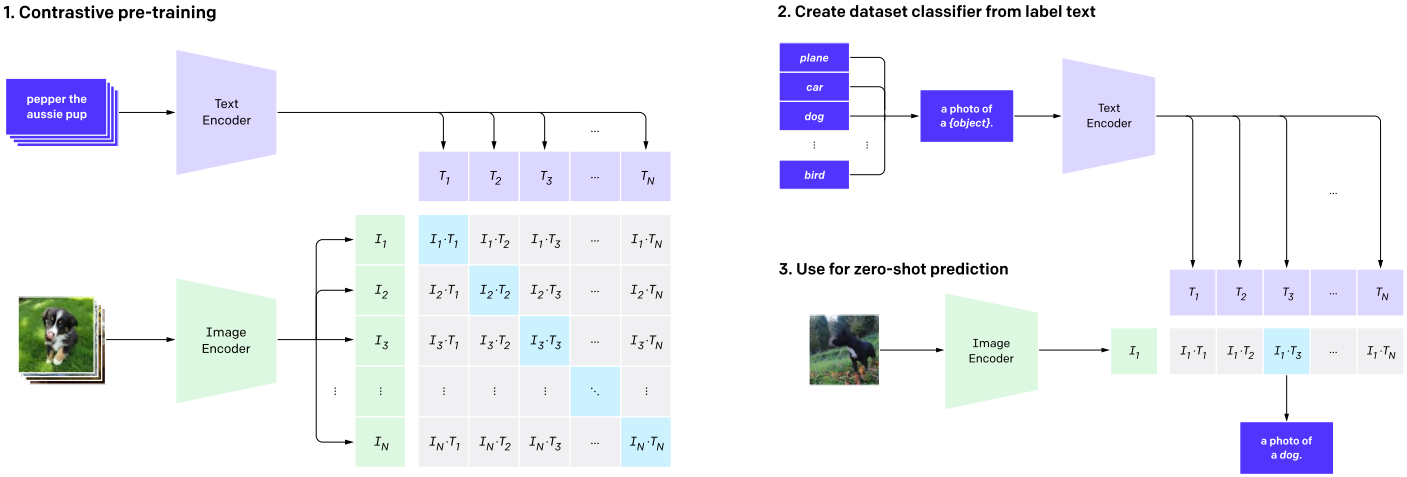
Existing approaches for response generation and dialogue state tracking struggle to handle text and picture simultaneously.

Datasets	Joint Accuracy	Inform Rate	Success Rate	BLEU Score	Combined Score	Image Match
WOZ 2.0	0.81	77.2	68.8	18.79	91.79	–
MultiWOZ 2.0	0.57	84.4	70.1	15.01	92.26	–
MultiWOZ 2.1	0.56	85.0	70.5	15.23	92.98	–
MMConv	0.28[2]	14.6[1]	9.2[1]	20.30	32.20	0.02

[1] We evaluate on predicted agent's action results. At least one exact venue should be correct to be count as informed.
[2] Here we exclude the effect of flexible open span here .

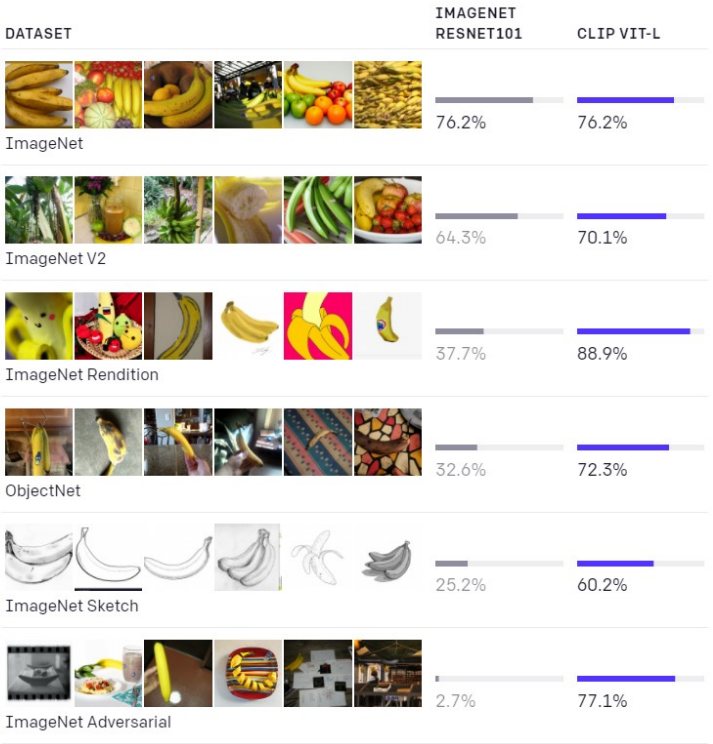
Running Demo

CLIP



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR.

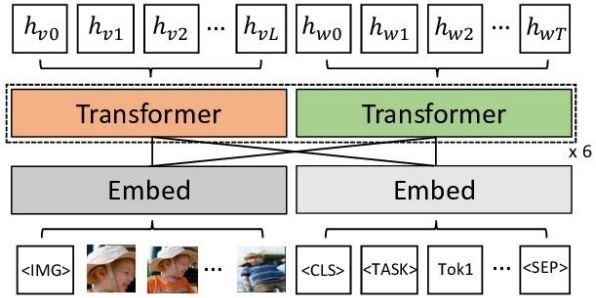
CLIP



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR.

ViLBERT

Examples of vision-and-language task



The overall architecture of ViLBERT.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*.

04
Future works

Ideas and things need to be done.

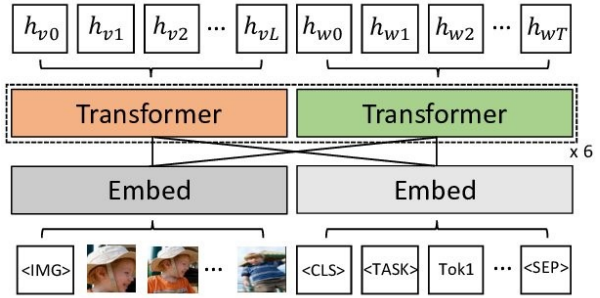
What's Next?

Evaluation

Automatic and human evaluation

Improve Model

Try to implement Dual Encoder



THANKS!

Do you have any questions?

Data
Science
Innovation
Lab.