Contrastively Pretrained Vision-Language Transformers and Domain Adaptation Methods for Multimodal TOD Systems

Youngjae Chang[1], Dooyoung Kim[1], Jinyoung Kim[1], Hyunmook Cha[1],
Keunha Kim[1], Sooyoung Min[1], Youngjoong Ko[1]*; Kye-Hwan Lee[2], Joonwoo Park[2]
Sungkyunkwan University[1], LG Electronics[2]
{youngjaechang0, kdysunleo98, jeankim941, chahyunmook, keunhakim98,
sujae9704}@gmail.com, yjko@skku.edu, {kyehwan.lee, joonwoo.park}@lge.com

Abstract

The Situated Interactive MultiModal Conversations (SIMMC2.1) Challenge 2022 is hosted by the Eleventh Dialog System Technology Challenge (DSTC11). The task of SIMMC is to create a shopping assistant agent that can communicate with customers in a virtual store. It requires processing store scenes and product catalogs along with the customer's request which could be decomposed into four steps and each becomes a subtask. In this work, we investigate monolithic transformers, fusion transformers, and language transformers as three distinct multimodal modeling approaches, and evaluate the potential of each. We also devise a retrieval-based method to acquire meta-data of each object which enhances the accuracy of predicted object characteristics significantly. Furthermore, we identify a discrepancy in using pretrained language models for dialog tasks and propose a simple domain-adaptation method. Our model came in third place for object coreferencing, dialog state tracking, and response generation tasks.

1 Introduction

Task-oriented dialog system refers to a system that communicates with users with a specific purpose, such as restaurant recommendation or accommodation reservation. The system should be able to analyze the user's utterance to track the dialog state and generate a response accordingly to fulfill the user's purpose. Recent advances in pretrained language models and new dialog datasets such as MultiWOZ have accelerated the research on task-oriented dialogs (Budzianowski et al., 2018).

To this end, Situated Interactive MultiModal Conversations (SIMMC) Challenge was proposed to introduce multimodality to the task-oriented dialog tasks (Crook et al., 2019). The challenge consists of a virtual shopping scenario in the fashion and furniture domain. Each dialog is based on up to two virtual scenes, and the user interacts with the system to purchase items in the corresponding domain. The system is required to process user utterances, previous dialogues, images of the scenes, and metadata files for the objects appearing in the scenes to find the adequate belief state and generate a response. An example of dialog is given in Figure 1, where the user seeks two pairs of trousers to purchase. As the dialogues are grounded on virtual scenes, the system should not only have a powerful language understanding but also should be able to process images and fuse the two modalities. In the Eleventh Dialog System Technology Challenge (DSTC11), SIMMC2.1 was held as the first track for competition. The SIMMC2.1 challenge decomposes the requirements of multimodal task-oriented systems into the following four subtasks: 1) Ambiguous Candidate Identification, 2) Multimodal Coreference Resolution, 3) Multimodal Dialog State Tracking, 4) Multimodal Dialog Response Generation. The only change from SIMMC2.0 (Kottur et al., 2021) is the first subtask, where the

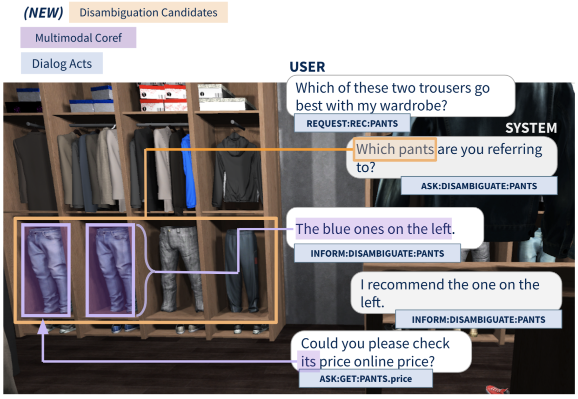
(NEW)
Disambiguation Candidates
Multimodal Coref
Dialog Acts
USER
Which of these two trousers go best with my wardrobe?
REQUEST:REC:PANTS
SYSTEM
Which pants are you referring to?
ASK:DISAMBIGUATE:PANTS
The blue ones on the left.
INFORM:DISAMBIGUATE:PANTS
I recommend the one on the left.
INFORM:DISAMBIGUATE:PANTS
Could you please check its price online price?
ASK:GET:PANTS.price

Figure 1: The SIMMC2.1 dataset consists of the dialogues between a user and a clerk, with an image of the virtual store's scene and the metadata for every object that appears in the scene. To achieve high performance on the given four subtasks, it is important to utilize all possible modalities with the most adequate method.

[*] Corresponding author

Proceedings of The Eleventh Dialog System Technology Challenge, pages 25–30
September 11, 2023 ©2023 Association for Computational Linguistics

binary classification task developed to be a more challenging candidate identification task.

Past studies have proposed various designs for solving the SIMMC task. For instance, Huang (Huang et al., 2021) proposed a model based on a vision-language transformer model and condensed various metadata and visual data of objects into a single token. Lee (Lee and Han, 2021) utilized contrastive learning to pretrain a dual encoder model to project image data and language data into the same semantic space. Son (Son et al., 2022) converted image data to a natural language form with multiple attribute classifiers and leveraged the power of a large pretrained language model for better language generation results. Likewise, approaches vary widely on how to process data from different modalities.

In this study, we identify three methods that are commonly used to model multimodality which is; monolithic vision language transformers, dual encoding vision language transformers, and language-centric multimodal transformers. We select a representative model from each category and conduct experiments on SIMMC2.1 to compare the performance of each method. We also develop a retrieval-oriented approach for obtaining the meta-data of individual objects, which substantially improves the precision of the predicted attributes for each object.

We find it unsuitable to fine-tune pretrained language models for the SIMMC task. Most pretrained language models pretrain on large text datasets that are crawled from the internet, which are predominantly web pages and documents. However, dialogues have different characteristics compared to typical text data because the vocabulary for colloquial style differs from written text (Zhang et al., 2020; Kulhánek et al., 2021). To be specific, spoken language tends to have frequent coreferencing and ellipsis, which require earlier contexts to understand. To this end, we experiment with several domain-adaptation methods based on pretraining to alleviate the problem.

2 Task Description

- **#1. Ambiguous Candidate Identification** The user utterance may be ambiguous depending on the given scene in the background. More than one item in the scene could match the user's description of an object, in which case there is the need to disambiguate the user's utterance. The goal of this subtask is to find all the ambiguous candidate objects in the scene that matches the user's description. As the objects are distinguished by canonical ID(s), the model should output all the canonical ID(s) of the ambiguous candidates.

- **#2. Multimodal Coreference Resolution** The system should know the object the user is referring to before preparing a response. To minimize error propagation, this step comes in high priority in the multimodal task-oriented dialog task. Similar to the first subtask, the model should output all the canonical ID(s) for the object the user is referring to.

- **#3. Multimodal Dialog State Tracking (MM-DST)** Task-oriented dialogues have a structured form of dialog states, also known as belief states. They are abstracts of the dialog and contain essential information for the system to fulfill the user's goals. This subtask involves predicting the correct user intent, and pairs of slot values the user provides.

- **#4. Multimodal Dialog Response Generation** The ultimate goal of the system is to generate an adequate answer to the user utterance. Utilizing the results from previous subtasks would give accurate guidance for generating a response. This task is evaluated by the similarity of the model output compared to the human response.

3 Methods

We present three representative methods for solving multimodal tasks and a pretraining method to enhance performance by adapting large pretrained language models to the dialog domain.

3.1 Monolithic Vision-Language Transformer

Previous studies have defined this category of vision-language transformers as monolithic because it tends to drastically simplify the processing of the visual inputs to be the same as the convolution-free manner in processing textual inputs (Kim et al., 2021). As depicted in <A> of Figure 2, we build object embeddings for every object in the scene by leveraging visual embeddings extracted from each object image, bounding box information, and textual embeddings extracted from the metadata. The information of different

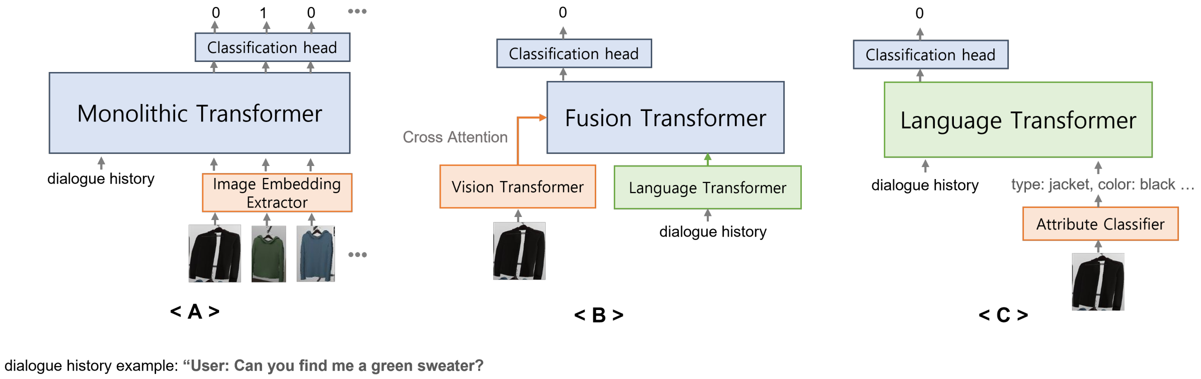

Figure 2: Recent studies have commonly used the three methods above for modeling multimodality. <A> model represents the Monolithic Vision-Language Transformer while <B> and <C> represent the Dual Vision-Language Transformer and the Language-Centric Multimodal Transformer, respectively.

modalities concatenated and projected by an MLP layer to match the hidden size of the monolithic vision-language transformer. All of the object embeddings are concatenated consecutively to the embeddings of dialog history and are processed by the monolithic vision-language transformer. Given a dialog, the dialog history will have consecutive turns, $(DH_i = t_1, t_2, ..., t_{i-1})$ where each turn as a user and system utterance $(t_i = u_i, s_i)$, and background scene with objects, $(S = o_1, o_2, ..., o_n)$, is created. Then, the input *x* for the model could be described as follows:

$$x_i = [DH_i \oplus u_i \oplus OA] \quad (1)$$

Where object attributes *OA* can be obtained by the following equation:

$$OA = MLP(VM(S) \oplus PLM(S) \oplus bbox) \quad (2)$$

The visual embedding extractor is a pretrained vision model *VM* and the textual embedding is a pretrained language model *PLM*.

3.2 Dual Vision-Language Transformer

We define vision-language models that use a separate encoder for modeling visual inputs and textual inputs as dual vision-language models. They are pretrained by a contrastive learning loss which induces the embeddings of visual inputs and image inputs to become similar if they contain the same semantics information (Lee and Han, 2021; Li et al., 2021). In <B> of Figure 2, a single object image and the dialog history are each processed by the vision transformer and the language transformer. The two output embeddings are fused by the cross-attention mechanism in the fusion transformer to be classified for final output. We use non-visual metadata as additional textual inputs for more coherence between vision and language encoders.

3.3 Language-Centric Multimodal Transformer

This method takes a pipeline approach to handle multimodal inputs, by converting data that are not language to a textual form. Recent studies have tested this method with various multimodal tasks such as visual question answering and yielded robust performance on those that generate natural language (Gao et al., 2022). We build an efficient input template for reflecting all the textual data for the language transformer to process. Each part is explained in detail as follows:

- **Object Metadata Prediction** We train the object data on a CNN model to extract attributes from the object images which can be obtained by utilizing the bounding box information and the scenes of the dialog data. We devise a retrieval-based method to reduce the search space of all possible attributes by training the image classifier to find the object itself among all the candidates in the catalog, instead of classifying each of the attributes. This way we drastically reduce the prediction space of each attribute and limit the model from outputting non-existent sets of attributes such as "pink check-patterned jeans". We convert the attributes extracted to a language form using a simple template for each attribute in the form of "[attribute type] is [predicted value].".

- **Language Transformer** As shown in <C> of Figure 2, we leverage the preprocessed objects and dialog history and perform classification on each of the objects with the output hidden representation of the language transformer. Instead of using the [CLS] token for classification, we utilized the soft-prompting technique by building an input template with extra learnable prompts. Prompt tuning (Lester et al., 2021) is simple to implement, yet efficient and explicitly changes the model behavior with additional input. Given that object description OD_j is the natural language form of attributes object o_j has, the following equation describes the template of input *x*.

$$x_i = [DH_i \oplus u_i \oplus [P1], [P2], [MASK] \oplus OD_j] \quad (3)$$

[P1], [P2] correspond to the two soft prompts we have used for this method, and the output of the [MASK] token is used for classification.

3.4 Domain-Adaptation Method

To mitigate the discrepancy between pretrained language models and dialog tasks, various methods of domain-adaptation have been proposed, and by far pretraining on a target domain has been most successful (He et al., 2022; Zhang et al., 2019). We conducted two different types of pretraining on the task dataset, which are random masked language modeling and noun prioritized masked language modeling. As we are pretraining an encoder-decoder model for generation, random masked language modeling works identically to the Token Masking of the BART reconstruction task. In the Token Masking task is masking a sequence randomly, and regenerate the masked sequence to its original form. This differs from fine-tuning because the objective is to learn general language representations by self-restoration. For the noun prioritized masked language modeling, we utilize the nltk POS tagger (Bird et al., 2009) to give additional weights to the nouns of the input for masking before applying the reconstruction loss. We presumed that this method would nudge the model to focus on the noun components of the input. This adjustment will mitigate the discrepancy problem because coreferencing and ellipsis usually occur on nouns in an utterance if they were mentioned earlier on in the dialogue. Given a sequence of tokens, $(T = t_1, t_2, ..., t_n)$, the corrupted version of the sequence, $(C = c_1, c_2, ..., c_m)$, the reconstruction loss can be described as follows:

$$\mathcal{L}_{RL} = -\sum_{i=1}^{|T|} \log p(t_i | C, t_1, t_2, ..., t_{i-1}; \theta) \quad (4)$$

Model	Disambiguation F1(%)	Coreferencing F1(%)
UNITER	57.0	67.5
ALBEF	**62.4**	**73.7**
RoBERTa-large	-	65.6

Table 1: Devtest evaluation of the three Representative methods of modeling multimodality. The disambiguation performance of RoBERTa-large is not evaluated due to its lack of performance on the easier task, coreferencing.

Name	DST		Response
	Intent F1(%)	Slot F1(%)	BLEU-4
BART	97.15	89.81	0.291
w/ pretrain	**97.35**	91.8	0.288
w/ noun priority	97.31	**92.1**	**0.297**

Table 2: Devtest evaluation of domain-adaptation methods.

4 Experiments

4.1 Experiment Settings

All three methods for modeling multimodality used a binary cross entropy loss for training and were evaluated on subtasks 1 and 2, where visual information plays a crucial role along with the input texts. For the monolithic vision-language transformer, we used the UNITER-based model (Huang et al., 2021) with visual embeddings extracted from our image preprocessor. In the case of the dual vision-language transformer, we leveraged ALBEF (Li et al., 2021) with a custom input template for the textual encoder. For the language-centric multimodal transformer, we set RoBERTa-large (Liu et al., 2019) as the backbone and utilized the ResNext101 model (Xie et al., 2017) as the attribute extractor.

For subtasks 3 and 4, we speculated that dialog state tracking and response generation rely heavily on the language understanding and generation power of the model. To address our speculation, we fine-tuned the BART model, a large pretrained language model that met our requirements. We experimented with the domain-adaptation methods by guiding BART with additional pretraining on our target dataset.

Team ID	Subtask 1	Subtask 2	Subtask 3		Subtask4
	F1	F1	Slot F1	Intent F1	BLEU-4
Team 1	0.6726	**0.9429**	**0.9424**	0.9598	**0.4093**
Team 2	0.6517	-	-	-	-
Team 3(Ours)	0.6384	0.7585	0.9048	0.9677	0.3029
Team 4	-	-	-	-	0.2519
Team 5	**0.705**	0.8028	0.9266	**0.9775**	0.365

Table 3: Final Results on the teststd split for SIMMC2.1 competition

4.2 Results

The results for the multimodal modeling methods are presented in Table 1, where ALBEF, the dual vision-language transformer yielded the highest performance followed by the monolithic vision-language transformer, UNITER. The language-centric multimodal model based on RoBERTa, had the worst performance. Given that the gap between the ALBEF and UNITER is significant, it is presumable that processing multiple objects at once may harm the performance of the system.

Performances on the domain-adaption method can be found in Table 2. The slot F1 score had a higher priority than the intent F1 score according to the rules of the competition. Under such criteria, the noun priority masking method for pretraining became the most preferable way for both dialog state tracking and response generation as it had higher performances in slot F1 and BLEU-4. In our analysis of specific cases of the model outputs, it was observed that the domain-adapted version demonstrates greater expressiveness, referencing past dialogues to employ a more comprehensive selection of adjectives for object elaboration. For instance, the baseline model provides a response with "I'll add that top to your cart", while the domain-adapted model produces a more expressive response, such as "I'll put that light grey tank top in your cart", which would in turn assist the user with a clearer understanding of the whole situation.

Table 3 presents the official results of SIMMC2.1 evaluated on the teststd data split. Before comparing our model with others, it should be acknowledged that we had different circumstances with teams 1 and 5. The two high-ranking teams have reused certain codes from a submission in SIMMC2.0 that might be problematic due to the changes in the allowed inputs during inference. The problem was about the same objects appearing in the train and test, which made it possible for the model to memorize all the attributes of the object by using the unique id and simple attribute classification heads. Disregarding this issue, we achieved 3rd place in subtasks 2, 3, and 4 with a marginal gap in performance for subtask 3.

5 Conclusion

In this work, we have identified 3 common ways to model multimodality and conducted experiments to find the best fit for multimodal task-oriented dialog systems. Despite that we found the dual vision-language transformer to have higher performance over other methods we did not set specific criteria for each method to be fairly evaluated such as the size of parameters and number of training steps. We also investigated domain-adaptation methods to enhance language understanding and generation in the dialog domain, where we found the noun-prioritized masking effective. It should be interesting to leverage part-of-speech tags for domain-adaptation to other domains, which could be material for future work.

6 Acknowledgements

This work was supported in part by LG Electronics (CTO/AI lab) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2C2100362).

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278.*

Paul A Crook, Shivani Poddar, Ankita De, Semir Shafi, David Whitney, Alborz Geramifard, and Rajen Subba. 2019. Simmc: Situated interactive multi-modal conversational data collection and evaluation platform. *arXiv preprint arXiv:1911.02690*.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

Yichen Huang, Yuchen Wang, and Yik-Cheung Tam. 2021. Uniter-based situated coreference resolution with rich multimodal input. *arXiv preprint arXiv:2112.03521*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*.

Jonáš Kulhánek, Vojtech Hudecek, Tomáš Nekvinda, and Ondrej Dušek. 2021. Augpt: Dialogue with pre-trained language models and data augmentation. *arXiv preprint arXiv:2102.05126*.

Joosung Lee and Kijong Han. 2021. Multimodal interactions using pretrained unimodal models for simmc 2.0. *arXiv preprint arXiv:2112.05328*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dongcheol Son, Youngjae Chang, Youngjoong Ko, and Jaehwan Lee. 2022. Domain Adaptive Task-Oriented Dialog System for Deep Understanding of Multimodal Conversation.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Yichi Zhang, Zhijian Ou, Huixin Wang, and Junlan Feng. 2020. A probabilistic end-to-end task-oriented dialog model with latent belief states towards semi-supervised learning. *arXiv preprint arXiv:2009.08115*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.