On the Effectiveness of Integration Methods for Multimodal Dialogue Response Retrieval

Seongbo Jang[1] **Seonghyeon Lee**[2] **Dongha Lee**[3] **Hwanjo Yu**[1,*]
[1]Pohang University of Science and Technology
[2]Kyungpook National University [3]Yonsei University
{jang.sb,hwanjoyu}@postech.ac.kr
sh0416@knu.ac.kr donalee@yonsei.ac.kr

Abstract

Multimodal chatbots have become one of the major topics for dialogue systems in both research community and industry. Recently, researchers have shed light on the multimodality of responses as well as dialogue contexts. This work explores how a dialogue system can output responses in various modalities such as text and image. To this end, we first formulate a multimodal dialogue response retrieval task for retrieval-based systems as the combination of three subtasks. We then propose three integration methods based on a two-step approach and an end-to-end approach, and compare the merits and demerits of each method. Experimental results on two datasets demonstrate that the end-to-end approach achieves comparable performance without an intermediate step in the two-step approach. In addition, a parameter sharing strategy not only reduces the number of parameters but also boosts performance by transferring knowledge across the subtasks and the modalities.

1 Introduction

As the demand for open-domain chatbots and technical development of dialogue systems rise steeply, researchers have brought attention to multimodal dialogue systems. Among various modalities, image-grounded conversations have been actively researched along with the advent of several benchmarks (Lin et al., 2014; Plummer et al., 2015; Antol et al., 2015; Das et al., 2017; Zellers et al., 2019). Most of the benchmarks focus on the factual contents of images, usually given in the form of question-answer pairs. In addition to factual information, recent studies started to consider emotional exchange and engagingness, which are humane aspects of open-domain dialogues (Hu et al., 2014), by collecting more chit-chat-like datasets such as image-grounded conversation (IGC) (Mostafazadeh et al., 2017) and ImageChat (Shuster et al., 2020a).

While most work focused on understanding dialogue contexts of multiple modalities, little work tried to build an integrated system that outputs multimodal responses for retrieval-based chatbots. Zang et al. (2021) proposed photo-sharing intent prediction and image retrieval as individual tasks for multimodal response retrieval, yet the combination of an image retriever and a text retriever remains ambiguous.

To overcome this limitation, we first formulate multimodal response retrieval task which aims to choose the most appropriate text or image response for the next utterance, given a dialogue context composed of text utterances. Then, we explore three unified methods that integrate subcomponents for the end task in different ways. To be specific, dual retriever (DR) and shared dual retriever (SDR) are based on a two-step approach: 1) intent prediction determines the modality of the next utterance and 2) response retrieval finds out the most likely utterance from the predicted modality; on the contrary, multimodal dual retriever (MDR) is an end-to-end approach that selects responses from a heterogeneous candidate pool of the both modalities without the explicit intent prediction step.

We evaluate the effectiveness of each method for multimodal response retrieval and investigate the effect of model size on two benchmark datasets. The two-step approach performs better than the end-to-end approach in unimodal retrieval tasks, whereas the end-to-end approach achieves comparable performance compared to the two-step approach in multimodal retrieval, posing a question on the necessity of intent prediction. In terms of model size, SDR and MDR reduce the number of parameters by sharing context encoders, which is impossible for DR that trains separate context encoders for each modality and intent prediction.

* Corresponding author

2 Related Work

2.1 Dialogue Response Retrieval

There are two representative approaches for dialogue response retrieval task: dual encoder and cross-encoder. Dual encoder approaches (Huang et al., 2013; Henderson et al., 2017; Mazaré et al., 2018; Dinan et al., 2019) encode dialogue contexts and responses separately with distinct encoders and compute matching scores to select the most appropriate response. This nature enables us to pre-compute representations to reduce computational burden at inference time, thus widely adopted in many studies (Humeau et al., 2020; Lan et al., 2021; Chen et al., 2021). In particular, researchers have focused on the selection of encoding functions such as CNNs (Yan et al., 2018; Wu et al., 2018), RNNs (Lowe et al., 2015), combined architectures of CNNs and RNNs (Yan et al., 2016; Zhou et al., 2016), memory networks (Zhang et al., 2018), transformers (Dinan et al., 2019; Xu et al., 2021b), and BERT (Henderson et al., 2019). Cross-encoder approaches (Gu et al., 2020; Whang et al., 2020; Wu et al., 2020; Xu et al., 2021a) take the concatenated inputs of contexts and responses to enrich interactions between the two. They are also actively studied owing to the development of pre-trained language models (PLMs) such as BERT (Devlin et al., 2019). In this work, we focus on dual encoder architecture since cross-encoders are less preferred to dual encoders in practical situations.

2.2 Multimodal Open-domain Dialogue

There have been several studies to build multimodal open-domain dialogue systems which output text responses related to a given image followed by a few turns of dialogue contexts. Shuster et al. (2020a) constructed ImageChat dataset which involves images and dialogue contexts along with personality traits allocated to speakers, and proposed a unified architecture using Transformer (Vaswani et al., 2017) and ResNet (He et al., 2016). Shuster et al. (2020b) built a multi-task dialogue agent using 12 open-domain dialogue datasets including ImageChat and IGC. On the success of Blender (Roller et al., 2021), Shuster et al. (2021) incorporated an image encoder (Xie et al., 2017) to enable image-grounded conversation. Sun et al. (2021) focused on generation models which generate either text or image responses conditioned on the preceding textual dialogue contexts. To the best of our knowledge, we propose a single integrated system for multimodal dialogue response retrieval for the first time.

3 Method

3.1 Task: Multimodal Response Retrieval

Multimodal response retrieval aims to select the most appropriate response among the text candidates and image candidates for a given dialogue context. The text is represented as a sequence of tokens $r^{\mathbf{t}} = (t_1, ..., t_l)$ where each token is included in a pre-defined vocabulary. The image is represented by 3-dimensional tensor $r^{\mathbf{i}} \in \mathbb{R}^{H \times W \times C}$. To solve this task, the model predicts the score $s(c, r)$ that indicates how much each response r in the candidate pool is appropriate for a dialogue context $c = (c_1, ..., c_l)$ where c_i is previous text utterance. Using this score, the model selects the response with the highest score for the given context. Therefore, the goal of the model is to accurately select the ground truth response while harmonizing the multimodal response candidates,

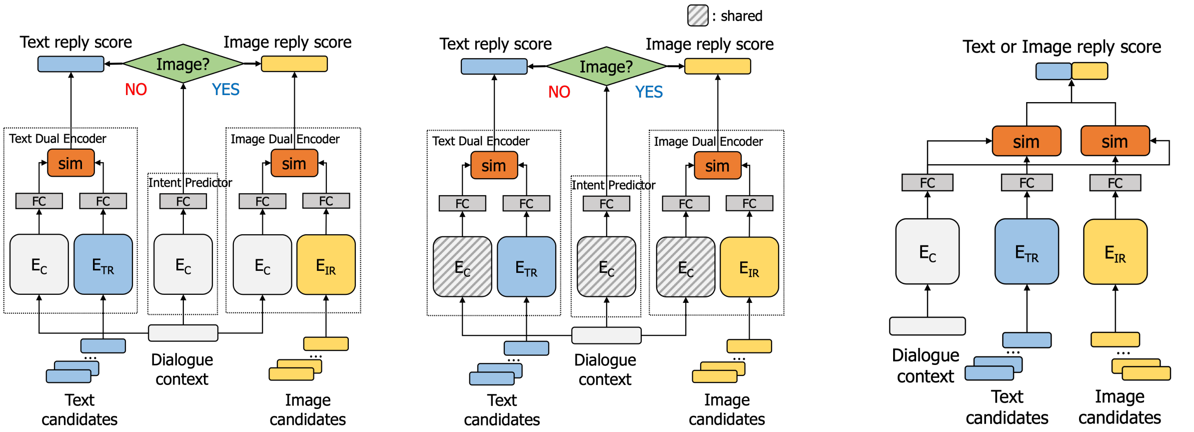
$$r^* = \underset{r \in \{r^{\mathbf{t}}_1, ..., r^{\mathbf{t}}_i\} \cup \{r^{\mathbf{i}}_1, ..., r^{\mathbf{i}}_j\}}{\operatorname{argmax}} s(c, r).$$

Subtask: Intent Prediction (Zang et al., 2021) aims to determine the modality of the next utterance given a context. In the case of two modalities (text and image), an intent prediction model f_i takes a context c as input and produces a binary logit. Given the pair of context and either image response or text response, the binary cross entropy loss is used for training as follows:

$$\mathcal{L}_{\text{intent}} = \textstyle\sum_{(c,r^i)} \mathcal{L}_{\text{BCE}}(f_i(c), 1) + \sum_{(c,r^t)} \mathcal{L}_{\text{BCE}}(f_i(c), 0).$$

Subtask: Text Response Retrieval is to select the most appropriate text response given the current context. We adopt dual encoder architecture widely used in response selection (Yang et al., 2018). To be specific, a context encoder $f^{\mathbf{t}}_c$ and a response encoder $f^{\mathbf{t}}_r$ compute the representations of a context c and a response r respectively. The cosine similarity between two representations is regarded as the score, and the encoders are optimized to accurately predict the score based on the cross entropy loss. The loss computed from i-th pair of context $c^{\mathbf{t}}_i$ and text response $r^{\mathbf{t}}_i$ is as follows:

$$\begin{aligned}\mathcal{L}_{\text{text}} = &- \log \frac{\exp(s_{\mathbf{t},\mathbf{t}}(c^{\mathbf{t}}_i, r^{\mathbf{t}}_i))}{\sum_{(\cdot, r^{\mathbf{t}}_j) \in B} \exp(s_{\mathbf{t},\mathbf{t}}(c^{\mathbf{t}}_i, r^{\mathbf{t}}_j))} \\ &- \log \frac{\exp(s_{\mathbf{t},\mathbf{t}}(c^{\mathbf{t}}_i, r^{\mathbf{t}}_i))}{\sum_{(c^{\mathbf{t}}_j, \cdot) \in B} \exp(s_{\mathbf{t},\mathbf{t}}(c^{\mathbf{t}}_j, r^{\mathbf{t}}_i))},\end{aligned}$$



(a) Dual Retriever (b) Shared Dual Retriever (c) Multimodal Dual Retriever

Figure 1: Three different architectures for the multimodal response retrieval task. E_C: context encoder, E_{TR}: text response encoder, E_{IR}: image response encoder.

where $s_{\cdot,\cdot}(c, r) = \cos(f_c^{\cdot}(c), f_r^{\cdot}(r))$ for further notational simplicity.

Subtask: Image Response Retrieval handles image responses during conversation. This task is the same with text response retrieval except that the modality of response is image. Therefore, the response encoder is built on the pretrained image encoder (He et al., 2016). The loss for the i-th pair of context $c_i^{\mathbf{t}}$ and image response $r_i^{\mathbf{i}}$ is described as follows:

$$\mathcal{L}_{\text{image}} = -\log \frac{\exp(s_{\mathbf{t},\mathbf{i}}(c_i^{\mathbf{t}}, r_i^{\mathbf{i}}))}{\sum_{(\cdot, r_j^{\mathbf{i}}) \in B} \exp(s_{\mathbf{t},\mathbf{i}}(c_i^{\mathbf{t}}, r_j^{\mathbf{i}}))} - \log \frac{\exp(s_{\mathbf{t},\mathbf{i}}(c_i^{\mathbf{t}}, r_i^{\mathbf{i}}))}{\sum_{(c_j^{\mathbf{t}}, \cdot) \in B} \exp(s_{\mathbf{t},\mathbf{i}}(c_j^{\mathbf{t}}, r_i^{\mathbf{i}}))}.$$

While the models for each subtasks are trained separately, these models do not capture useful supervision across the subtasks and cannot effectively solve the multimodal response retrieval task. Therefore, careful integration of these models become important for the ultimate task. Considering the knowledge derived across the subtasks, we introduce three approaches (Fig. 1) to integrate these modules in the following subsections.

3.2 Dual Retriever (DR)

One simple integration is to weave the separately trained models. Each model is trained by three different optimization problems:

$$\underset{\theta_i}{\text{minimize}}\ \mathcal{L}_{\text{intent}},\ \underset{\theta_c^{\mathbf{t}}, \theta_r^{\mathbf{t}}}{\text{minimize}}\ \mathcal{L}_{\text{text}},\ \underset{\theta_c^{\mathbf{t}}, \theta_r^{\mathbf{i}}}{\text{minimize}}\ \mathcal{L}_{\text{image}},$$

where θ represents the parameters of each model. Using these trained models, DR combines the produced outputs to obtain final outputs during inference. To be specific, the intent predictor predicts the modality of the response for an input context. Depending on the prediction, we select the corresponding retrieval model and then find out the most appropriate response.

$$f(c^{\mathbf{t}}) = \begin{cases} \text{argmax}_{r_i^{\mathbf{i}}}\ s_{\mathbf{t},\mathbf{i}}(c^{\mathbf{t}}, r_i^{\mathbf{i}}) & \text{if } f_i(c) > 0.5 \\ \text{argmax}_{r_i^{\mathbf{t}}}\ s_{\mathbf{t},\mathbf{t}}(c^{\mathbf{t}}, r_i^{\mathbf{t}}) & \text{otherwise.} \end{cases}$$

However, since the supervision from each modality is not transferred across different retrievers, the derived knowledge is not fully reflected to both retrieval models.

3.3 Shared Dual Retriever (SDR)

We propose a simple but effective scheme that shares the retriever to encourage the models to communicate with each other across the subtasks. Our key idea comes from the observation that the two retrieval tasks follow the same dual encoder architecture. Also, although the architecture of response encoder is different due to the different modality, the architecture of context encoder is the same. Thus, we can share the parameters of the context encoder between the two subtasks without any modification of the architecture ($\theta_c^{\mathbf{t}} = \theta_c^{\mathbf{t}}$). With the help of parameter sharing, we integrate the optimization problems for two response retrieval tasks:

$$\underset{\theta_i}{\text{minimize}}\ \mathcal{L}_{\text{intent}},\ \underset{\theta_c^{\mathbf{t}}, \theta_r^{\mathbf{t}}, \theta_r^{\mathbf{i}}}{\text{minimize}}\ \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{image}}.$$

Furthermore, we make the context encoder inside the intent predictor ($\theta_i = \theta_c$) share its parameters with those of the response retrieval models. By doing so, the separate optimization problems are merged into a unified optimization problem for the multimodal dialogue response retrieval task as follows:

$$\underset{\theta_c^{\mathbf{t}},\theta_r^{\mathbf{t}},\theta_r^{\mathbf{i}}}{\text{minimize}}\ \mathcal{L}_{\text{intent}} + \mathcal{L}_{\text{image}} + \mathcal{L}_{\text{text}}.$$

We further hypothesize the ineffectiveness inside the inference process due to the intent predictor. One reason is the cascaded error coming from the intent predictor. The intent predictor acts as a branch for choosing the modality of the most appropriate response. However, it means that the final prediction result is wrong if the intent predictor predicts wrong modality. In addition, from the recent success in modeling cross-modal representation space (Radford et al., 2021; Akbari et al., 2021), we hypothesize that the response representation space from different modalities can become naturally aligned.

3.4 Multimodal Dual Retriever (MDR)

From the above hypothesis, we propose our final integration approach by removing the intent predictor and modeling multimodal response representation space. We remove the intent predictor and directly compare the cosine similarities across different modality. Then an integrated response encoder is defined as $f_r^{\mathbf{m}} = f_r^{\mathbf{i}}$ if $r = r^{\mathbf{i}}$, or $f_r^{\mathbf{m}} = f_r^{\mathbf{t}}$ if $r = r^{\mathbf{t}}$. The loss for the i-th pair of the two different modalities is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{joint}} = &- \log \frac{\exp(s_{\mathbf{t},\mathbf{m}}(c_i^{\mathbf{t}}, r_i))}{\sum_{(\cdot,r_j)\in B} \exp(s_{\mathbf{t},\mathbf{m}}(c_i^{\mathbf{t}}, r_j))} \\ &- \log \frac{\exp(s_{\mathbf{t},\mathbf{m}}(c_i^{\mathbf{t}}, r_i))}{\sum_{(c_j,\cdot)\in B} \exp(s_{\mathbf{t},\mathbf{m}}(c_j^{\mathbf{t}}, r_i))},\end{aligned}$$

where a batch B consists of context-response pairs whose response is either image or text. All the parameters can be effectively optimized to minimize the joint loss in an end-to-end manner,

$$\underset{\theta_c,\theta_{r^{\mathbf{t}}},\theta_{r^{\mathbf{i}}}}{\text{minimize}}\ \mathcal{L}_{\text{joint}}.$$

Note that the absence of the intent predictor further simplifies the inference process by taking the response that has the most highest score among all the multimodal candidates.

4 Experimental Setup

4.1 Datasets

PhotoChat is a multimodal dialogue dataset which includes dyadic dialogues covering various topics in daily lives. It consists of 10,286/1,000/1,000 dialogue contexts with a single image attached to each context in the train/dev/test set. There exist 8,889/1,000/1,000 unique images along with object labels in the train/dev/test set.

Multi-modal Dialogue (MMDial) (Lee et al., 2021) was constructed by substituting text utterances in existing text-only dialogue datasets with relevant images from large-scale image datasets using a state-of-the-art image-text matching model (Li et al., 2019). It consists of 39,956/2,401/2,673 dialogue contexts with a single image attached to each context in the train/dev/test set. There exist 12,272/334/682 unique images in the train/dev/test set.

Preprocessing. For each dialogue context, we extract the first $n \in \{1, 2, \ldots, l\}$ utterances as an example to augment contexts, where l is the length of a context before sharing an image. Since object labels are not explicitly attached in MMDial, we match the corresponding image captions to images from MS-COCO and Flickr30k.

4.2 Metric

We measure recall at k (R@k) to evaluate multimodal response retrieval. It calculates the occurrence when the gold response is retrieved in top-k ($k = 1, 5, 10$) candidates. For dev and test, we use the fixed 50 candidates of each modality randomly chosen from the whole images and text responses in the dev and the test set, respectively. We use the checkpoint that achieves the best dev R@k and report the test R@k evaluated by the checkpoint.

5 Results

5.1 Effects of Integration Methods

In contrast to relatively small gaps of R@k among the three methods in unimodal retrieval, the end task performance is largely affected depending on how we integrate the subcomponents. On PhotoChat (Table 1), SDR achieves the highest R@k for multimodal retrieval, except that MDR reaches slightly higher R@10 when the model size is large. Meanwhile on MMDial (Table 2), MDR achieves the highest R@k for multimodal retrieval, except

		Text Retrieval			Image Retrieval			Multimodal Retrieval		
Model Size	Method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Small	DR	**0.4325**	0.7371	0.8701	**0.4519**	**0.7994**	**0.9007**	0.2475	0.4234	0.4949
	SDR	0.4274	**0.7553**	**0.8721**	0.4396	0.7787	0.8873	**0.4000**	**0.7046**	**0.8086**
	MDR	0.3990	0.7076	0.8518	0.4122	0.7269	0.8599	0.3766	0.6594	0.7721
Large	DR	0.4305	0.7482	0.8751	**0.4620**	**0.8247**	**0.9058**	0.2079	0.3560	0.4143
	SDR	**0.4650**	**0.7990**	**0.9066**	0.4315	0.8061	0.9046	**0.4152**	**0.7239**	0.8157
	MDR	0.4315	0.7299	0.8812	0.4406	0.7797	0.8964	0.4020	0.6949	**0.8193**

Table 1: Results on the test set of PhotoChat. Text (image) retrieval: retrieval among 50 text (image) response candidates for the examples whose ground truths are text (image) responses, assuming the intent predictor is an oracle for DR and SDR. MDR does not get advantage in this setting since there is no explicit intent prediction step. Small: $\text{BERT}_{\text{MINI}}$ and ResNet_{50}, Large: $\text{BERT}_{\text{BASE}}$ and ResNet_{152} (Turc et al., 2019).

		Text Retrieval			Image Retrieval			Multimodal Retrieval		
Model Size	Method	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Small	DR	0.6118	0.8039	**0.8946**	0.0430	0.2148	0.3687	**0.2874**	0.3749	0.4161
	SDR	**0.6138**	**0.8115**	0.8870	**0.1356**	**0.3958**	**0.5430**	0.2619	0.3934	0.4527
	MDR	0.5386	0.7558	0.8512	0.1281	0.3846	0.5334	0.2731	**0.4517**	**0.5378**
Large	DR	**0.6333**	**0.8453**	**0.9141**	0.0605	0.2204	0.3544	0.2900	0.3835	0.4125
	SDR	0.5986	0.8313	0.8986	**0.1484**	**0.4300**	**0.6002**	0.2516	0.4113	0.4735
	MDR	0.5967	0.8202	0.8938	0.1281	0.4097	0.5728	**0.3079**	**0.4883**	**0.5895**

Table 2: Results on the test set of MMDial. Defined terms are the same as in Table 1.

that DR scores the highest R@1 for the small model. We note that although DR outperforms MDR in text retrieval and image retrieval, MDR outperforms DR on the contrary due to the cascaded error from the intent prediction step. These results highlight the significance of choosing an appropriate integration method of submodules.

5.2 Effects of Model Size

Overall, all the methods in large models tend to achieve higher R@k than their counterparts in small models on the both datasets. On PhotoChat, MDR shows comparable performance for multimodal retrieval to SDR when the model size grows large. Similarly on MMDial, MDR effectively increases R@k for multimodal retrieval compared to the two-step approach when the large model is used. From these results, we can conclude that large models are more capable of aligning the multimodal representation space than small models.

5.3 Effects of Parameter Sharing

On the both datasets, the performance of DR for multimodal retrieval lags behind those of SDR and MDR which share the parameters of context encoders. DR trains each submodule separately on the three individual subtasks so none of the subcomponents can get the knowledge from the other subtasks and modalities positively transferred, resulting in disharmony for accomplishing the ultimate goal (Wu et al., 2021).

In addition, weight sharing decreases the number of total parameters from 72M to 49M in small models and from 501M to 281M in large models, which become around 1.5x and 1.8x smaller respectively.

6 Conclusion

We propose an integrated task to build a multimodal dialogue system that outputs both text and image responses, and present three architectures for the task, named DR, SDR, and MDR, respectively. We then analyze their advantages and disadvantages in terms of retrieval performance and model size.

Specifically, we empirically analyze the effectiveness of intent prediction which was introduced in the previous work. The experimental results on two datasets demonstrate that the end-to-end approach without an intent predictor shows competitiveness for multimodal retrieval compared to the two-step approach. In addition, SDR and MDR successfully reduce the number of model parameters without compromising the end task performance. As the end-to-end approach is the efficient method to solve the end task directly, elaborating cross-modal interactions in this architecture will become one promising direction for our future work.

References

Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Wei Chen, Yeyun Gong, Can Xu, Huang Hu, Bolun Yao, Zhongyu Wei, Zhihao Fan, Xiaowu Hu, Bartuer Zhou, Biao Cheng, et al. 2021. Contextual fine-to-coarse distillation for coarse-grained response selection in open-domain conversations. *arXiv preprint arXiv:2109.13087*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404, Florence, Italy. Association for Computational Linguistics.

Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What we instagram: A first analysis of instagram photo content and user types. In *Eighth International AAAI conference on weblogs and social media*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Tian Lan, Deng Cai, Yan Wang, Yixuan Su, Xian-Ling Mao, and Heyan Huang. 2021. Exploring dense retrieval for dialogue response selection. *arXiv preprint arXiv:2110.06612*.

Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyon Myaeng. 2021. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 897–906, Online. Association for Computational Linguistics.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural

Language Processing, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

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 Jason Weston. 2020b. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470, Online. Association for Computational Linguistics.

Kurt Shuster, Eric Michael Smith, Da Ju, and Jason Weston. 2021. Multi-modal open-domain dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4863–4883, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2021. Multimodal dialogue response generation. *arXiv preprint arXiv:2110.08515*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. In *INTERSPEECH*, pages 1585–1589.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Ruihan Wu, Chuan Guo, Awni Hannun, and Laurens van der Maaten. 2021. Fixes that fail: Self-defeating improvements in machine-learning systems. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Yu Wu, Zhoujun Li, Wei Wu, and Ming Zhou. 2018. Response selection with topic clues for retrieval-based chatbots. *Neurocomputing*, 316:251–261.

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.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021a. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14158–14166.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021b. Topic-aware multi-turn dialogue modeling. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64.

Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, and Zhoujun Li. 2018. Response selection from unstructured documents for human-computer conversation systems. *Knowledge-Based Systems*, 142:149–159.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. PhotoChat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

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.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381, Austin, Texas. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

Architectural details of encoders The context encoder (E_C) and the text response encoder (E_{TR}) consist of a single BERT encoder followed by a projection layer. The parameters of two encoders are shared across all experiments since this yields better performance consistently than when not shared. The image response encoder (E_{IR}) consists of an image encoder which extracts visual features with ResNet and a object label encoder which extracts object label features with BERT. Note that both BERT and ResNet are used to encode image features since an object label is attached to every image. The two representations are then concatenated and projected to the joint embedding space. We use $\text{BERT}_{\text{MINI}}$ and ResNet_{50} for small models and $\text{BERT}_{\text{BASE}}$ and ResNet_{152} for large models as specified by Turc et al. (2019).

Hyperparameters For BERT, the dropout rate is 0.2 and the maximum sequence length is set to 128. We use cosine similarity with the temperature $\tau = 0.01$ as the similarity measure. We train the model for 10 epochs for intent prediction and text text dialogue retrieval, and 20 epochs for image dialogue retrieval and multimodal dialogue retrieval with the batch size of 64 in one V100 GPU. We apply Adam optimizer (Kingma and Ba, 2015) with the learning rate of 5e-5 and linear decay of 0.1% per 1,000 steps.