Improving Context Modelling in Multimodal Dialogue Generation

Shubham Agarwal,[*] **Ondřej Dušek, Ioannis Konstas and Verena Rieser**

The Interaction Lab, Department of Computer Science
Heriot-Watt University, Edinburgh, UK
[*]Adeptmind Scholar, Adeptmind Inc., Toronto, Canada
{sa201, o.dusek, i.konstas, v.t.rieser}@hw.ac.uk

Abstract

In this work, we investigate the task of textual response generation in a multimodal task-oriented dialogue system. Our work is based on the recently released Multimodal Dialogue (MMD) dataset (Saha et al., 2017) in the fashion domain. We introduce a multimodal extension to the Hierarchical Recurrent Encoder-Decoder (HRED) model and show that this extension outperforms strong baselines in terms of text-based similarity metrics. We also showcase the shortcomings of current vision and language models by performing an error analysis on our system's output.

1 Introduction

This work aims to learn strategies for textual response generation in a multimodal conversation directly from data. Conversational AI has great potential for online retail: It greatly enhances user experience and in turn directly affects user retention (Chai et al., 2000), especially if the interaction is multi-modal in nature. So far, most conversational agents are uni-modal – ranging from open-domain conversation (Ram et al., 2018; Papaioannou et al., 2017; Fang et al., 2017) to task oriented dialogue systems (Rieser and Lemon, 2010, 2011; Young et al., 2013; Singh et al., 2000; Wen et al., 2016). While recent progress in deep learning has unified research at the intersection of vision and language, the availability of open-source multimodal dialogue datasets still remains a bottleneck.

This research makes use of a recently released Multimodal Dialogue (MMD) dataset (Saha et al., 2017), which contains multiple dialogue sessions in the fashion domain. The MMD dataset provides an interesting new challenge, combining recent efforts on task-oriented dialogue systems, as well as visually grounded dialogue. In contrast to simple QA tasks in visually grounded dialogue, e.g. (Antol et al., 2015), it contains conversations with a clear end-goal. However, in contrast to previous slot-filling dialogue systems, e.g. (Rieser and Lemon, 2011; Young et al., 2013), it heavily relies on the extra visual modality to drive the conversation forward (see Figure 1).

In the following, we propose a fully data-driven response generation model for this task. Our work is able to ground the system's textual response with language and images by learning the semantic correspondence between them while modelling long-term dialogue context.

Figure 1: Example of a user-agent interaction in the fashion domain. In this work, we are interested in the textual response generation for a user query. Both user query and agent response can be multimodal in nature.

2 Model: Multimodal HRED over multiple images

Our model is an extension of the recently introduced Hierarchical Recurrent Encoder Decoder (HRED) architecture (Serban et al., 2016, 2017;

129

Lu et al., 2016). In contrast to standard sequence-to-sequence models (Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015), HREDs model the dialogue context by introducing a context Recurrent Neural Network (RNN) over the encoder RNN, thus forming a hierarchical encoder.

We build on top of the HRED architecture to include multimodality over multiple images. A simple HRED consists of three RNN modules: encoder, context and decoder. In multimodal HRED, we combine the output representations from the utterance encoder with concatenated multiple image representations and pass them as input to the context encoder (see Figure 2). A dialogue is modelled as a sequence of utterances (turns), which in turn are modelled as sequences of words and images. Formally, a dialogue is generated according to the following:

$$P_\theta(t_1, \ldots t_N) = \prod_{n=1}^{N} P_\theta(t_n | t_{<n}) \qquad (1)$$

where t_n is the n-th utterance in a dialogue. For each $m = 1, \ldots, M_n$, we have hidden states of each module defined as:

$$h_{n,m}^{text} = f_\theta^{text}(h_{n,m-1}^{text}, w_{m,n}) \qquad (2)$$

$$h_n^{img} = l^{img}([g_\theta^{enc}(img_1), \ldots g_\theta^{enc}(img_k)]) \qquad (3)$$

$$h_n^{cxt} = f_\theta^{cxt}(h_{n-1}^{cxt}, [h_{n,M_n}^{text}, h_n^{img}]) \qquad (4)$$

$$h_{n,m}^{dec} = f_\theta^{dec}(h_{n,m-1}^{dec}, w_{n,m}, h_{n-1}^{cxt}) \qquad (5)$$

$$h_{n,0}^{text} = 0; \quad h_0^{cxt} = 0; \quad h_{n,0}^{dec} = h_N^{cxt} \qquad (6)$$

where $f_\theta^{text}, f_\theta^{cxt}$ and f_θ^{dec} are GRU cells (Cho et al., 2014). θ represent model parameters, $w_{n,m}$ is the m-th word in the n-th utterance and g_θ^{enc} is a Convolutional Neural Network (CNN); here we use VGGnet (Simonyan and Zisserman, 2014). We pass multiple images in a context through the CNN in order to get encoded image representations $g_\theta^{enc}(img_k)$. Then these are combined together and passed through a linear layer l^{img} to get the aggregated image representation for one turn of context, denoted by h_n^{img} above. The textual representation h_{n,M_n}^{text} is given by the encoder RNN f_θ^{text}. Both h_{n,M_n}^{text} and h_n^{img} are subsequently concatenated and passed as input to the context RNN. h_N^{cxt}, the final hidden state of the context RNN, acts as the initial hidden state of the decoder RNN. Finally, output is generated by passing $h_{n,m}^{dec}$ through an affine transformation followed by a softmax activation. The model is trained using cross entropy on next-word prediction. During generation, the decoder conditions on the previous output token.

Saha et al. (2017) propose a similar baseline model for the MMD dataset, extending HREDs to include the visual modality. However, for simplicity's sake, they 'unroll' multiple images in a single utterance to include only one image per utterance. While computationally leaner, this approach ultimately loses the objective of capturing multi-modality over the context of multiple images and text. In contrast, we combine all the image representations in the utterance using a linear layer. We argue that modelling all images is necessary to answer questions that address previous agent responses. For example in Figure 3, when the user asks "what about the 4th image?", it is impossible to give a correct response without reasoning over all images in the previous response. In the following, we empirically show that our extension leads to better results in terms of text-based similarity measures, as well as quality of generated dialogues.

Our version of the dataset
Text Context: Sorry i don't think i have any 100 % acrylic but i can show you in knit | Show me something similar to the 4th image but with the material different
Image Context: [Img 1, Img 2, Img 3, Img 4, Img 5] | [0, 0, 0, 0, 0]
Target Response: The similar looking ones are
Saha et al. (Saha et al., 2017)
Text Context: |
Image Context: Img 4 | Img 5
Target Response: The similar looking ones are

Figure 3: Example contexts for a given system utterance; note the difference in our approach from Saha et al. (2017) when extracting the training data from the original chat logs. For simplicity, in this illustration we consider a context size of 2 previous utterances. '\|' differentiates turns for a given context. We concatenate the representation vector of all images in one turn of a dialogue to form the image context. If there is no image in the utterance, we consider a 0_{4096} vector to form the image context. In this work, we focus only on the textual response of the agent.
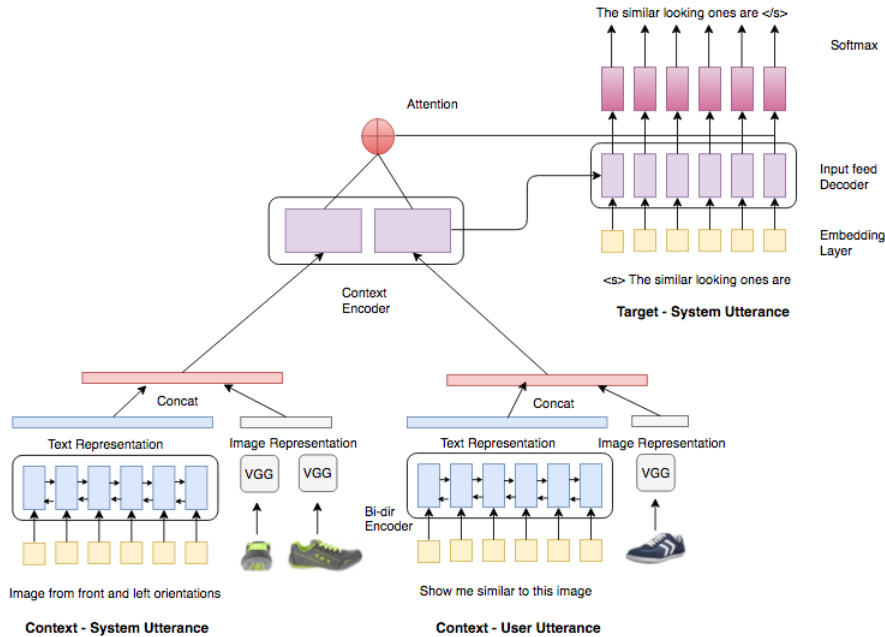
130

Figure 2: The Multimodal HRED architecture consists of four modules: utterance encoder, image encoder, context encoder and decoder. While Saha et al. (2017) 'rollout' images to encode only one image per context, we concatenate all the 'local' representations to form a 'global' image representation per turn. Next, we concatenate the encoded text representation and finally everything gets fed to the context encoder.

3   Experiments and Results

3.1   Dataset

The MMD dataset (Saha et al., 2017) consists of 100/11/11k train/validation/test chat sessions comprising 3.5M context-response pairs for the model. Each session contains an average of 40 dialogue turns (average of 8 words per textual response, 4 images per image response). The data contains complex user queries, which pose new challenges for multimodal, task-based dialogue, such as *quantitative inference (sorting, counting and filtering)*: "Show me more images of the 3rd product in some different directions", *inference using domain knowledge and long term context*: "Will the 5th result go well with a large sized messenger bag?", *inference over aggregate of images:* "List more in the upper material of the 5th image and style as the 3rd and the 5th", *co-reference resolution*. Note that we started with the raw transcripts of dialogue sessions to create our own version of the dataset for the model. This is done since the authors originally consider each image as a different context, while we consider all the images in a single turn as one concatenated context (cf. Figure 3).

3.2   Implementation

We use the PyTorch[1] framework (Paszke et al., 2017) for our implementation.[2] We used 512 as the word embedding size as well as hidden dimension for all the RNNs using GRUs (Cho et al., 2014) with tied embeddings for the (bidirectional) encoder and decoder. The decoder uses Luong-style attention mechanism (Luong et al., 2015) with input feeding. We trained our model with the Adam optimizer (Kingma and Ba, 2015), with a learning rate of 0.0004 and clipping gradient norm over 5. We perform early stopping by monitoring validation loss. For image representations, we use the FC6 layer representations of the VGG-19 (Simonyan and Zisserman, 2014), pre-trained on ImageNet.[3]

3.3   Analysis and Results

We report sentence-level BLEU-4 (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and ROUGE-L (Lin and Och, 2004) using the evaluation scripts provided by (Sharma et al., 2017).

---

[1] https://pytorch.org/

[2] Our code is freely available at: https://github.com/shubhamagarwal92/mmd

[3] In future, we plan to exploit state-of-the-art frameworks such as ResNet or DenseNet and fine tune the image encoder jointly, during the training of the model.
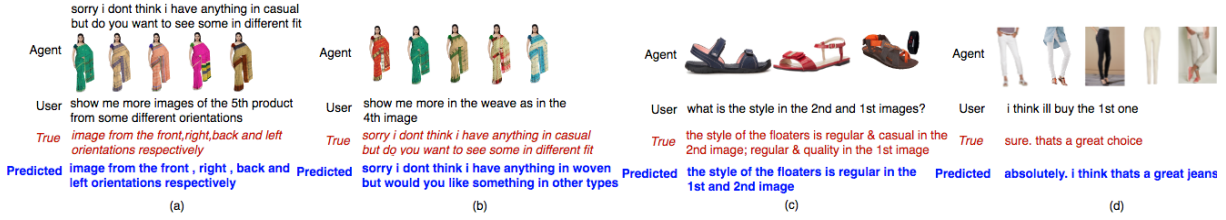
Figure 4: Examples of predictions using M-HRED–attn (5). Recall, we are focusing on generating textual responses. Our model predictions are shown in blue while the true gold target in red. We are showing only the previous user utterance for brevity's sake.

We compare our results against Saha et al. (2017) by using their code and data-generation scripts.[4] Note that the results reported in their paper are on a different version of the corpus, hence not directly comparable.

Model	Cxt	BLEU-4	METEOR	ROUGE-L
Saha et al. M-HRED*	2	0.3767	0.2847	0.6235
T-HRED	2	0.4292	0.3269	0.6692
M-HRED	2	0.4308	0.3288	0.6700
T-HRED–attn	2	0.4331	0.3298	0.6710
M-HRED–attn	2	0.4345	0.3315	0.6712
T-HRED–attn	5	0.4442	**0.3374**	0.6797
M-HRED–attn	5	**0.4451**	0.3371	**0.6799**

Table 1: Sentence-level BLEU-4, METEOR and ROUGE-L results for the response generation task on the MMD corpus. "Cxt" represents context size considered by the model. Our best performing model is M-HRED–attn over a context of 5 turns. *Saha et al. has been trained on a different version of the dataset.

Table 1 provides results for different configurations of our model ("T" stands for text-only in the encoder, "M" for multimodal, and "attn" for using attention in the decoder). We experimented with different context sizes and found that output quality improved with increased context size (models with 5-turn context perform better than those with a 2-turn context), confirming the observation by Serban et al. (2016, 2017).[5] Using attention clearly helps: even T-HRED–attn outperforms M-HRED (without attention) for the same context size. We also tested whether multimodal input has an impact on the generated outputs. However, there was only a slight increase in BLEU score (M-HRED–attn vs T-HRED–attn).

To summarize, our best performing model (M-HRED–attn) outperforms the model of Saha et al. by 7 BLEU points.[6] This can be primarily attributed to the way we created the input for our model from raw chat logs, as well as incorporating more information during decoding via attention. Figure 4 provides example output utterances using M-HRED–attn with a context size of 5. Our model is able to accurately map the response to previous textual context turns as shown in (a) and (c). In (c), it is able to capture that the user is asking about the style in the 1st and 2nd image. (d) shows an example where our model is able to relate that the corresponding product is 'jeans' from visual features, while it is not able to model fine-grained details like in (b) that the style is 'casual fit' but resorts to 'woven'.

4 Conclusion and Future Work

In this research, we address the novel task of response generation in search-based multimodal dialogue by learning from the recently released Multimodal Dialogue (MMD) dataset (Saha et al., 2017). We introduce a novel extension to the Hierarchical Recurrent Encoder-Decoder (HRED) model (Serban et al., 2016) and show that our implementation significantly outperforms the model of Saha et al. (2017) by modelling the full multimodal context. Contrary to their results, our generation outputs improved by adding attention and increasing context size. However, we also show that multimodal HRED does not improve significantly over text-only HRED, similar to observations by Agrawal et al. (2016) and Qian et al. (2018). Our model learns to handle textual correspondence between the questions and answers, while mostly ignoring the visual context. This indicates that we need better visual models to en-

---

[4] https://github.com/amritasaha1812/MMD_Code

[5] Using pairwise bootstrap resampling test (Koehn, 2004), we confirmed that the difference of M-HRED-attn (5) vs. M-HRED-attn (2) is statistically significant at 95% confidence level.

[6] The difference is statistically significant at 95% confidence level according to the pairwise bootstrap resampling test (Koehn, 2004).

code the image representations when he have multiple similar-looking images, e.g., black hats in Figure 3. We believe that the results should improve with a jointly trained or fine-tuned CNN for generating the image representations, which we plan to implement in future work.

Acknowledgments

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *Proceedings of EMNLP*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of ICCV*, pages 2425–2433.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Joyce Yue Chai, Nanda Kambhatla, and Wlodek Zadrozny. 2000. Natural language sales assistant-a web-based dialog system for online sales. In *Proceedings of AAAI*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*.

Hao Fang, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mari Ostendorf, Yejin Choi, and Noah A Smith. 2017. Sounding board–university of washington's alexa prize submission. *Alexa Prize Proceedings*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR abs/1412.6980*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of 2nd Workshop on Statistical Machine Translation*, pages 228–231.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of ACL*, pages 605–612.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Proceedings of NIPS*, pages 289–297.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *Proceedings of EMNLP*.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondrej Dušek, Verena Rieser, and Oliver Lemon. 2017. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. *Alexa Prize Proceedings*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Xin Qian, Ziyi Zhong, and Jieli Zhou. 2018. Multimodal machine translation with reinforcement learning. *CoRR abs/1805.02356*.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational AI: The science behind the alexa prize. *CoRR abs/1801.03604*.

Verena Rieser and Oliver Lemon. 2010. Natural language generation as planning under uncertainty for spoken dialogue systems. In *Empirical methods in natural language generation*, pages 105–120. Springer.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer.

Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2017. Multimodal dialogs (MMD): A large-scale dataset for studying multimodal domain-aware conversations. *CoRR abs/1704.00200*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of AAAI*, pages 3295–3301.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR abs/1706.09799*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556*.

Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. 2000. Reinforcement learning for spoken dialogue systems. In *Proceedings of NIPS*, pages 956–962.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, pages 3104–3112.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *CoRR abs/1604.04562*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.