A Multi-view Meta-learning Approach for Multi-modal Response Generation

Zhiliang Tian
National University of Defense Technology
Changsha, China
tianzhiliang@nudt.edu.cn

Zheng Xie
National University of Defense Technology
Changsha, China
xiezheng81@nudt.edu.cn

Fuqiang Lin
National University of Defense Technology
Changsha, China
linfuqiang13@alumni.nudt.edu.cn

Yiping Song*
National University of Defense Technology
Changsha, China
songyiping@nudt.edu.cn

ABSTRACT

As massive conversation examples are easily accessible on the Internet, we are now able to organize large-scale conversation corpora to build chatbots in a data-driven manner. Multi-modal social chatbots produce conversational utterances according to both textual utterances and vision signals. Due to the difficulty of bridging different modalities, the dialogue generation model of chatbots falls into local minima that only capture the mapping between textual input and textual output, as a result, it almost ignores the non-textual signals. Further, similar to the dialogue model with plain text as input and output, the generated responses from multi-modal dialogue also lack diversity and informativeness. In this paper, to address the above issues, we propose a Multi-View Meta-Learning (MultiVML) algorithm that groups samples in multiple views and customizes generation models to different groups. We employ a multi-view clustering to group the training samples so as to attend more to the unique information in non-textual modality. Tailoring different sets of model parameters for each group boosts the genereation diversity via meta-learning. We evaluate MultiVML on two variants of the OpenViDial benchmark datasets. The experiments show that our model not only better explore the information from multiple modalities, but also excels baselines in both quality and diversity.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**.

KEYWORDS

Multi-modal, Response generation, Meta-learning

ACM Reference Format:
Zhiliang Tian, Zheng Xie, Fuqiang Lin, and Yiping Song. 2023. A Multi-view Meta-learning Approach for Multi-modal Response Generation. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543507.3583548

*Corresponding Author

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04...$15.00
https://doi.org/10.1145/3543507.3583548

1 INTRODUCTION

Along with the maturity of Web 2.0, there has been an explosion in the number of people having conversations on websites with chatbots on social media (e.g., Facebook, Twitter). Most chatbots lead a text-to-text conversation between chatbots and users [52]. However, people prefer face-to-face communication due to the accessibility of the speaker's visual signals, e.g., facial expressions. Recently, building multi-modal chatbots is becoming a hot research topic, where the multi-modal generation model behind the chatbots [37, 65] can comprehend both textual and non-textual signals.

Neural generation models are widely-used in text-to-text response generation, but in multi-modal response generation, they suffer from some issues when additionally handling the corresponding non-textual signals. First, researchers find that most multi-modal text generation models rely highly on the textual input and attend less to the non-textual signals (i.e. vision signals) [1], which is called the "modality-bias" problem [16, 32, 45, 53, 66]. The reason is the models find the input-output mapping between the same modalities (i.e. text-to-text) is easier to capture; thus, the model falls into the local minima that are mainly based on the text-to-text mapping. Some researchers address this issue by enhancing the multi-modal information fusion using attention mechanism [5, 60, 66]. But those works are good at incorporating non-textual information which is related to the textual input, but still tend to ignore the non-overlapped information between different modalities. Second, multi-modal response generation also inherits the disadvantages of the plain text-to-text generation model: the generated outputs lack diversity and informativeness [27]. The reason is that the models tend to capture only the most salient input-output mapping since it is trained via maximizing likelihood estimation (MLE) [22, 27, 72]. Hence, some researchers enhance the generation diversity by refining the training objective [25, 56], incorporating randomness in model training [75], and importing additional information [18, 48, 73]. These methods need either a quality-diversity trade-off or additional resources.

Facing the above two issues, we argue that (1) multi-modal text generation should cater more to the non-overlapped information

[1] According to our experiment (see details in Sec.5.7), for a Transformer-based generation model that feeds the image feature as one time step in the encoder, 97.61% attention of the decoder locates on textual input and 2.39% locates on visual input.