Emotion-Aware Multimodal Pre-training for Image-Grounded Emotional Response Generation

Zhiliang Tian[1], Zhihua Wen[2], Zhenghao Wu[1], Yiping Song[3], Jintao Tang[3], Dongsheng Li[2(✉)], and Nevin L. Zhang[1]

[1] The Hong Kong University of Science and Technology, Sai Kung, Hong Kong SAR, China
{ztianac,lzhang}@cse.ust.hk, zwubq@connect.ust.hk

[2] Science and Technology on Parallel and Distributed Laboratory, National University of Defense Technology, Changsha, Hunan, China
{zhwen,dsli}@nudt.edu.cn

[3] National University of Defense Technology, Changsha, Hunan, China
{songyiping,tangjintao}@nudt.edu.cn

Abstract. Face-to-face communication leads to better interactions between speakers than text-to-text conversations since the speakers can capture both textual and visual signals. Image-grounded emotional response generation (IgERG) tasks requires chatbots to generate a response with the understanding of both textual contexts and speakers' emotions in visual signals. Pre-training models enhance many NLP and CV tasks and image-text pre-training also helps multimodal tasks. However, existing image-text pre-training methods typically pre-train on images by recognizing or modeling objects, but ignore the emotions expressed in the images. In this paper, we propose several pre-training tasks in a unified framework that not only captures emotions from images but also learns to incorporate the emotion into text generation. The pre-training involves single-modal learning to strengthen the ability to understand images and generate texts. It also involves cross-modal learning to enhance interactions between images and texts. The experiments verify our method in appropriateness, informativeness, and emotion consistency.

Keywords: Multimodal · Conversation · Emotion · Pre-training · Generation

1 Introduction

Most conversation systems [21,45,59] lead a text-to-text dialog between users and chatbots. However, most people prefer face-to-face communication due to the accessibility of the speaker's visual signals, like facial expressions and body language. After analysing those signals, chatbots can garner speakers' emotional

Z. Tian and Z. Wen–The two authors contributed equally to this work.

A. Bhattacharya et al. (Eds.): DASFAA 2022, LNCS 13247, pp. 3–19, 2022.
https://doi.org/10.1007/978-3-031-00129-1_1