

Perceive before Respond: Improving Sticker Response Selection by Emotion Distillation and Hard Mining

Supplementary Material

Anonymous Author(s)
Submission Id: 2084

In the supplementary material, we provide the descriptions of the main notations, explore different emotion knowledge distillation methods, design a simple framework to validate the generality of the proposed training paradigm and introduce more visualization results. The code for our model is in the attachment.

1 THE MAIN NOTATIONS

The main notations in the manuscript and the corresponding descriptions are shown in Table. 1.

2 DETAILS OF THE DATASETS

In this section, we provide details of the two datasets used in the manuscript, *i.e.*, StickerChat [2], DSTC10-MOD [1] and SER30K [6].

StickerChat is a multi-modal multi-turn dialog dataset, collected from public chat groups in messaging apps. The 20 utterances before each sticker image are stored as the historical context of the conversation, and together with the sticker image form a sample pair. It consists of 320,168 training pairs, 10,000 validation pairs, and 10,000 testing pairs. Each sentence in the training set contains an average of 7.54 words and each dialogue contains an average of 5.81 users. StickerChat contains a total of 174,695 sticker images across 3516 topics, and the average number of stickers in a topic is 49.64. Furthermore, the author of the stickers assigns each sticker with a corresponding emoji tag to convey its semantic or emotional meaning.

DSTC10-MOD is a competition dataset published by WeChat Conversation Platform, which consists of real open-domain conversations and is available in both Chinese and English. Following [7], we only adopt the Chinese version of DSTC10-MOD. It contains 45,000 open-domain dialogs with 307 stickers, each sample consists of a history dialog and stickers appearing in the dialog. Since the competition has ended, and to our knowledge the test set is currently no longer available, all experiments in the main manuscript are evaluated on the validation set.

SER30K is a large-scale sticker emotion recognition dataset, its sticker images are crawled from a sticker image website and annotated by three expert annotators. The dataset comprises of 1,887 sticker topics with a total of 30,739 sticker images, with training, validation, and test sets accounting for 70%, 10%, and 20% respectively. Each sticker is labeled with one of the seven emotion categories, *i.e.*, sadness, disgust, surprise, happiness, fear, anger and neutral. In addition, about 19% of the images in the dataset contain textual content, which consists of commonly used phrases in daily conversations, resulting in relatively shorter text length (most samples have text lengths lower than 6).

3 DETAILS OF THE TEACHER MODEL

We use the standard ResNet50 [3] as the teacher model for the EKD module. The batch size is set to 128, training a total of 100 epochs on SER30K [6]. The input image is scaled to 128×128 and then randomly rotated and randomly horizontally flipped. We apply the Adam optimizer with a learning rate of 10^{-4} , adjusting its decay schedule to decrease the learning rate by a factor of 0.1 at the 50th and 80th epochs. The final classification accuracy of the teacher model on the SER30K test set is 64.56%, which can be considered as an upper bound on the performance of the student model.

4 VISUALIZATION RESULTS

Fig. 1 shows visualization results in the StickerChat [2] validation set, where (a) and (b) are examples of common usages of stickers, (c) and (d) are two representative failure cases. The main usage of Stickers in dialogues is Strategic (*i.e.*, maintenance of social status quo, forming sympathy) and Functional (*e.g.*, substitute for text, supplement for text) [4, 5]. In (a), the user uses the sticker to express a different emotion from the dialogue. The sticker in this example mainly plays a strategic role, forming sympathy among users in the conversation. The user expresses his attitude toward the story by “hahaha”, which means he finds the story interesting. But at the end of the dialogue, he responds with a sticker of a cat cowering under the quilt, and the textual content in the sticker is *weak, pathetic, and helpless*. The user complements his sentiment with such a sticker and expresses sympathy for his little niece who lost teeth. In (b), the user uses the sticker to supplement the text. In the historical dialogue, the user expresses annoyance that the old mouse is not damaged when he wants to buy a new one. He then responds with a sticker of a weeping cartoon character to emphasize his negative emotion.

In (c) and (d) we show two failure cases. We show the Top-6 stickers with the model prediction scores, where the score of the ground truth sticker is shown with a green background. Some stickers contain textual content, which is closely related to the semantics and emotions expressed by these stickers. For example, the ground truth sticker shown in (c) contains textual content, which must be taken into consideration in order to make an accurate prediction. However, we did not construct a corresponding branch to analyze the textual content in the sticker. As a result, the model may mis-rank the candidate stickers even if the top-ranked sticker seems to match the conversation. We believe that introducing optical character recognition can improve the understanding of stickers by the model, which can be a future research direction. Some sticker responses are related to the user’s personal preferences. We show in (d) that multiple suitable sticker responses could exist in the candidate set. Since the StickerChat dataset is collected in real conversation scenarios, the

sticker responses are related to the user’s characteristics. Although the sticker with the first ranking score predicted by our model also matches the conversation, it does not match the user’s preference. We believe that investigating the modeling of personalized sticker responses for individual users could be a valuable future research direction.

REFERENCES

- [1] Zhengcong Fei, Zekang Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. Towards expressive communication with internet memes: A new multimodal conversation dataset and benchmark. *arXiv preprint arXiv:2109.01839* (2021).
- [2] Shen Gao, Xiuying Chen, Chang Liu, Li Liu, Dongyan Zhao, and Rui Yan. 2020. Learning to respond with stickers: A framework of unifying multi-modality in multi-turn dialog. In *WWW*.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
- [4] Artie Konrad, Susan C Herring, and David Choi. 2020. Sticker and emoji use in facebook Messenger: Implications for graphicon change. *Journal of Computer-Mediated Communication* 25 (2020), 217–235.
- [5] Joon Young Lee, Nahi Hong, Soomin Kim, Jonghwan Oh, and Joonhwan Lee. 2016. Smiley face: why we use emoticon stickers in mobile messaging. In *MobileHCI*.
- [6] Shengzhe Liu, Xin Zhang, and Jufeng Yang. 2022. SER30K: A Large-Scale Dataset for Sticker Emotion Recognition. In *ACM MM*.
- [7] Zhixin Zhang, Yeshuang Zhu, Zhengcong Fei, Jinchao Zhang, and Jie Zhou. 2022. Selecting stickers in open-domain dialogue through multitask learning. In *ACL*.

Table 1: Summary of the main notations in the manuscript as well as the corresponding descriptions. Note that all the notations have been explained in the manuscript.

Notation	Description
U	the input multi-turn dialogue
u_i	the i -th utterance in the dialogue
N_U	the number of utterances in the dialogue
N_S	the number of candidate sticker images
S	the candidate sticker images
s_i	the i -th image in the candidate sticker images
f	the ranking model
pos	the index of the ground truth sticker in S
neg	the negative samples in dialogue-sticker matching
H, W	the height and width of the input image
V	the vision features of image encoder output
v_i	the i -th vision feature of the image encoder output
v_{cls}	the global vision feature
N_V	the length of sequence vision features
C_V	the dimension of vision features
W	the text representations
w_i	the i -th word embedding in W
N_T	the number of words
C_T	the dimension of word embedding
w_{cls}	the average of all word features
K	the number of samples in a mini-batch
\hat{V}, \hat{W}	the features of vision and language modalities in a common embedding space
\hat{v}, \hat{w}	the [CLS] token embeddings of \hat{V} and \hat{W}
P	the probability predicted by multimodal encoder
C_S	the dimensions of SER30K image features extracted by the teacher model
M	the number of clusters for each emotion category in the EKD module
E	the Emotion Anchor feature matrix in the EKD module
V^e	the StickerChat image features extracted by the teacher model
V_{aug1}, V_{aug2}	the output of the image encoder for two parallel data augmentations
v_{aug1}, v_{aug2}	the [CLS] token embeddings of each sticker
P^T, N^T	the positive and negative score of the dialogue
α, β, γ	the trade-offs between the individual objective functions

