Supplementary Material

Abshishek Rajora*, Shubham Gupta*, Suman Kundu
Indian Institute of Technology Jodhpur, India
{rajora.1,gupta.37,suman}@iitj.ac.in

1. Dataset & Settings

CrisisMMD [2] is widely recognised multimodal crisis dataset. This dataset extracted from the social platform named as Twitter, comprises images and textual content related to seven major disasters from 2017, including hurricanes, earthquakes, floods, and wildfires. It is divided into three distinct tasks: Task 1 aims to classify image-text pairs as informative or non-informative. Task 2 focuses on categorizing the impact of the event into five classes: vehicle damage, infrastructure damage, affected individuals (injuries, fatalities, missing, found, etc.), rescue efforts, and others. Task 3 is dedicated to assessing severity, categorizing it as severe, mild, or little/no damage. We conduct our evaluations of these tasks under two distinct setups, akin to the state-of-the-art (SOTA) method described in [1]. In Setting A, our model is trained exclusively on image-text pairs that share identical labels, ensuring a consistent label match between pairs during training. Conversely, Setting B is more inclusive, allowing for the training on all types of labeled image-text pairs, regardless of label consistency. However, the testing data in Setting B is aligned to mirror the conditions of Setting A, to maintain comparability in evaluation metrics. The statistics of train, validation, and test data set for each task and setting is detailed in Table 1, providing a clear view of how the data is segmented and utilized in each experimental framework.

Table 1. Dataset distribution across various split, tasks and settings.

Setting	Task	Train	Val	Test	Total
A	Task 1	9601	1573	1534	12708
	Task 2	2874	477	451	3802
	Task 3	2461	529	530	3520
B	Task 1	13608	1573	1534	16715
	Task 2	8348	477	451	9276

2. Experimental Setup & Evaluation Metrics

We configured our model training across various tasks and settings using parameters viz. base learning rate of 2×10^{-3}, decay rate of 10X, batch size of 8, Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-4}$, $\lambda_1 = 0.4$ & $\lambda_2 = 0.6$, and total epochs of 50. Parameters for the graph and text transformers are consistent with those specified in the original studies [3, 4]. For the masking encoder, we have retained the model with same parameter combinations as used in [6] along with encoder depth of 32, image size of 228, and patch size of 14. These models are run on NVIDIA A30 with 24 GB of GPU memory. As part of textual pre-processing, symbols such as '@', '#' and hyperlinks are removed. The model's performance is assessed using metrics such as classification accuracy, macro F1-score, weighted F1-score, and MTMS (Multi-task Model Strength) [5].

*Both authors contribute equally.

References

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689, 2020. 1

[2] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018. 1

[3] Deng Cai and Wai Lam. Graph transformer for graph-to-sequence learning. In *AAAI*, pages 7464–7471. AAAI Press, 2020. 1

[4] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*. OpenReview.net, 2020. 1

[5] Shubham Gupta, Nandini Saini, Suman Kundu, and Debasis Das. Crisiskan: Knowledge-infused and explainable multimodal attention network for crisis event classification. In *Advances in Information Retrieval*, pages 18–33, Cham, 2024. Springer Nature Switzerland. 1

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 1