Differential Attention for Multimodal Crisis Event Analysis

Nusrat Munia[1], Junfeng Zhu[1], Olfa Nasraoui[2], and Abdullah-Al-Zubaer Imran[1]
[1]University of Kentucky, [2]University of Louisville

Code

Problem

Crisis responders drown in millions of mixed image–text posts and struggle to spot actionable information fast

Contribution

- **VLM-Enhanced Knowledge Fusion**: Leverages vision–language models to improve image-text alignment.
- **Encoder Boost**: Strengthens features using frozen pretrained VLMs, with zero additional training.
- **Fusion Strategies**: Compare multiple fusion approaches to align vision-and-text representations.
- **SOTA on CrisisMMD** [1]: Consistently surpasses prior multimodal techniques across all three benchmark tasks.

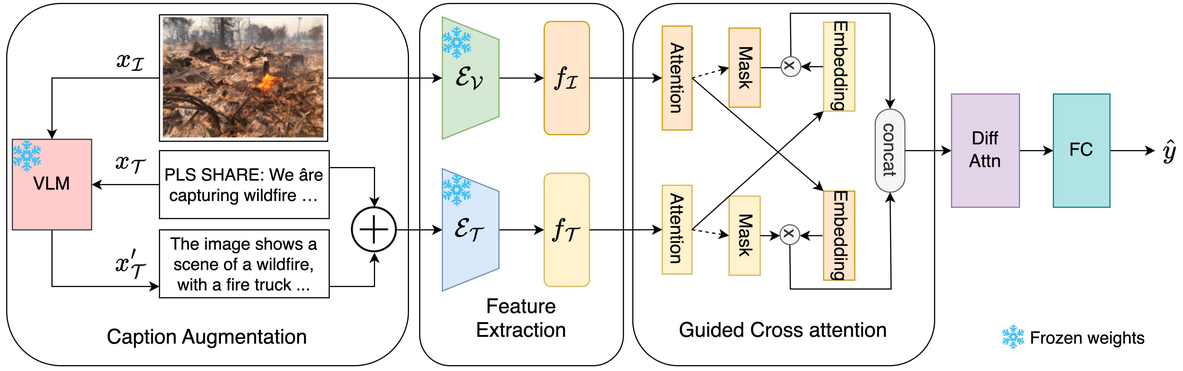
CrisisMMD Dataset

Task	Classes	Train	Val	Test
Informativeness	Informative	6,343	1,056	1,030
	Not-informative	3256	517	504
Humanitarian	Infrastructure Damage	595	78	78
	Vehicle Damage	17	2	2
	Rescue Efforts	912	149	126
	Affected Individuals	47	8	7
	Others	1,279	239	235
Damage Severity	Severe	1,548	332	332
	Mild	587	126	126
	Little Or No Damage	333	71	71

Acknowledgements

This work is supported by the National Science Foundation under Cooperative Agreement No. 2344533.

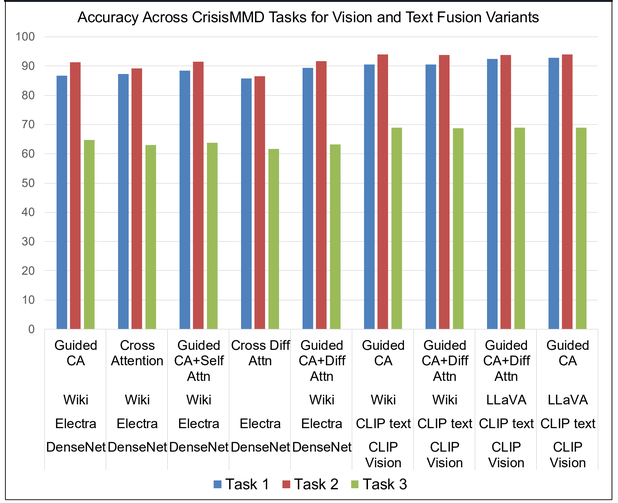
Method



- **Caption Augmentation**: Utilize LLaVA [2] to generate captions x'_T for images and combine with text x_T.
- **Feature Extraction**: Obtain frozen CLIP [3] embeddings f_I and f_T from images and enriched text.
- **Guided Cross Attention(CA)** [4]: Compute modality specific attention masks and fuse features via concatenation.
- **Differential Attention** [5]: Apply differential attention to emphasize complementary modality interactions.
- **Classification head**: Use fully connected layer for final classification.

Image	Tweet	LLaVA generated Text
	Residents recall the night they had to flee the \#redwoodvalleyfire.	The image shows a woman standing in a destroyed, burned-out building, surrounded by debris and rubble. The scene appears to be a result of a fire that has ravaged the area, leaving behind a desolate and destroyed landscape. The woman is likely a resident who had to flee the fire and has returned to the scene to assess the damage or to search for any belongings that might have been lost during the fire. The image captures the aftermath of a disaster, highlighting the destruction and the challenges faced by those affected by such events.
	Harvey, Irma could ding U.S. economy for combined \$290 billion #finance	The image shows a street filled with debris, including fallen trees and branches, and a street sign. There is also a truck and a person riding a bicycle in the scene. The street appears to be damaged due to a storm, possibly a hurricane, as suggested by the caption. The storm has caused significant damage to the area, including fallen trees and branches, which could potentially impact the local economy. The combined cost of the damage could be around \$290 billion, as indicated by the tweet. The image captures the aftermath of the storm, highlighting the need for recovery efforts and the potential economic consequences of such natural disasters.

Results



Conclusion

- LLaVA-generated captions add a rich, visual description to every crisis image.
- Frozen CLIP encoders supply well-aligned embeddings straight out of the box.
- Guided Cross-Attention with Differential Attention enables adaptive fusion between image and text modalities.

References

1. Alam, F. et al. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. ICWSM, 2018.
2. Liu, H. et. al. Visual instruction tuning. Advances in Neural Information Processing Systems, 2023.
3. Radford, A. et al. (2021). Learning transferable visual models from natural language supervision, ICML, 2021.
4. Gupta, S. et. al. Crisiskan: Knowledge-infused and explainable multimodal attention network for crisis event classification. ECIR, 2024.
5. Ye, T. et al. Differential Transformer. ICLR, 2025.