Supplementary Materials: Multimodal Categorization of Crisis Events in Social Media

Mahdi Abavisani
Dataminr Inc., New York, NY
mabavisani@dataminr.com

Liwei Wu
Department of Statistics
University of California, Davis
Davis, CA
liwu@ucdavis.edu

Shengli Hu
Dataminr Inc., New York, NY
shu@dataminr.com

Joel Tetreault
Dataminr Inc., New York, NY
jtetreault@dataminr.com

Alejandro Jaimes
Dataminr Inc., New York, NY
ajaimes@dataminr.com

Abstract

In previous experiments of this paper, we followed prior research in crisis event categorization and viewed the task as a multi-class single-label task. In this section, we provide three simple modifications to our model for extending it to a multi-label multi-class classifier.

1. Setting D Multi-Label Multi-class Categorization

In a multimodal single-label classification system, representations of different modalities are often fused to construct a joint representation from which a common label is reasoned for the multimodal-pair. Our classifiers in settings A, B, and C are multimodal multi-class single-label models. However, in setting D, we are interested in using both image and text information to predict separate labels for them. Figure 1 (a) and (b) show examples of these settings.

In Figure 1 (a), the multimodal pair, including image and text are both labeled as *Vehicle Damage*. On the contrary, in Figure 1 (b), while the image shows damaged vehicles, the text-only contains information about the location of the event and therefore does not fall in the *Vehicle Damage* category. In setting D, we want to use the information in both image and text to classify the image of this example into the *Vehicle Damage* class and the text into the *Other Relevant Information* class.

Cross-Attention: A straightforward way to capture these properties is by attaching two classifier heads to the output of the cross-attention module in our proposed model. We refer to this version as *Cross-Attention* classifier.

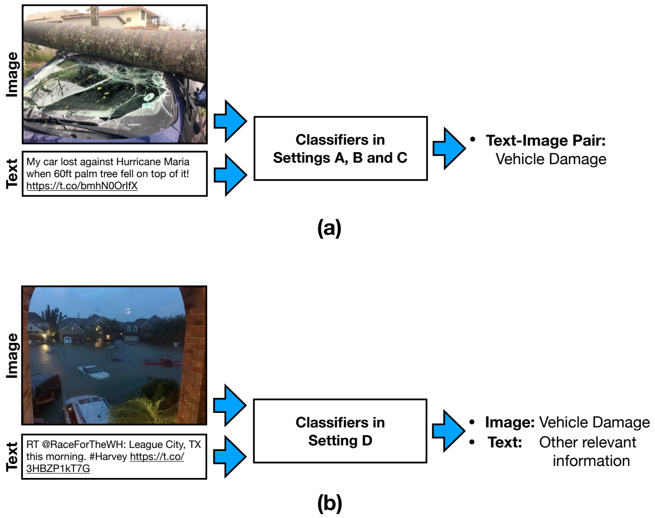
Image
Text
Classifiers in Settings A, B and C
• Text-Image Pair: Vehicle Damage
(a)
Image
Text
Classifiers in Setting D
• Image: Vehicle Damage
• Text: Other relevant information
(b)

My car lost against Hurricane Maria when 60ft palm tree fell on top of it! https://t.co/bmhN0OrlfX

RT @RaceForTheWH: League City, TX this morning. #Harvey https://t.co/3HBZP1kT7G

Figure 1. The behavior of our classifiers in different settings. (a) Our classifiers in settings A, B, and C view the task as a multi-class single-label task. (b) Our classifiers in setting D view the task as a multi-class multi-label task.

Self-Attention: The cross-attention mechanism in Eq. (4) uses text embeddings (image feature maps) to block misleading information from image feature maps (text embeddings). However, in setting D, since image and text may have different labels, they both can be informative but contain different information. Thus, we replace this module by separate self-attention blocks [2, 4] in each modality. That is, we still filter the uninformative features, but we do that based on the information in the modality itself.

Self-Cross-Attention: In the *Self-Attention* extension, the features of different modalities do not interact directly with each other. With a few modifications to the self-attention extension and combining it with our cross-attention model, one can develop a version of our method that is specifically designed for multi-label multi-class classification tasks. We use a self-attention block to learn a mask that filters the uninformative features from the modalities. In the meantime, we invert this mask and use the invert mask to attend to the other modality for selecting useful features. This way, not only do we develop modality-specific features, but we do so by exploiting useful information from both modalities. Let γ_{v_i} and γ_{t_i} be the self-attention masks that are calculated as:

$$\begin{aligned}\gamma_{v_i} &= \sigma(W_v''^T[f_i] + b_v''),\\ \gamma_{e_i} &= \sigma(W_e''^T[e_i] + b_e'')\end{aligned} \tag{1}$$

From equation (1), we can calculate the inverse-masks by

$$\begin{aligned}\gamma'_{v_i} &= 1 - \gamma_{v_i}\\ \gamma'_{e_i} &= 1 - \gamma_{e_i}.\end{aligned} \tag{2}$$

After we have the attention masks and the inverse of them, we can calculate the augmented image features f''_i and augmented text feature e''_i as

$$\begin{aligned}f''_i &= \gamma v_i \cdot \tilde{f}_i + \gamma'_{v_i} \cdot \tilde{e}_i\\ e''_i &= \gamma t_i \cdot \tilde{e}_i + \gamma'_{t_i} \cdot \tilde{f}_i\end{aligned} \tag{3}$$

where $\tilde{e}_i$ and $\tilde{f}_i$ are same as in Eq. (3) in the paper. We feed f''_i and e''_i to classifier heads of images and texts, respectively.

1.1. Experiments:

We evaluate the multi-label extensions in Task 1. In this experiment, both training and test sets contain inconsistent labels. That is in both training and testing we may have:

$$\mathcal{C}(v_i) \neq \mathcal{C}(t_i), \tag{4}$$

As the test set of this setting contains samples with inconsistent labels for image and text, we set $0 < p_0^t < 1$ for the training cases so that we include inconsistent image-text labels in training as well. In particular, we use $\Phi_t = \{p_0^t : 0.27, \rho_t : 900\}$ and $\Phi_v = \{p_0^v : 0.36, \rho_v : 900\}$. Benchmarks for this setting include unimodal models as well as a version of the feature fusion model with two classification heads.

We evaluate our method on Task 1. We keep the ratio between the number of samples in train and test sets similar to setting B in Table 2. However, we randomly sample with relaxing the Eq. (9) assumption of the paper for both the train and test sets.

Table 1. Setting D: Informativeness Evaluation

Model		Acc	Macro F1	Weighted F1
DenseNet [3]	Images :	78.30	78.30	78.31
BERT [1]	Text :	82.63	74.93	80.87
Feature Fusion	Images :	78.37	78.15	78.21
	Texts:	83.63	79.01	83.22
Cross-Attention	Images :	77.17	77.51	77.51
	Texts:	83.35	**79.60**	**83.41**
Self-Attention	Images :	**82.56**	**82.54**	**82.56**
	Texts:	**83.63**	76.79	82.17
Self-Cross-Attention	Images :	81.64	81.51	81.55
	Texts:	83.45	78.22	82.78

In Table 1, the result of different methods are compared in terms of Accuracy, Macro-F1, and Weighted F1. By comparing unimodal DenseNet and BERT results with Table 4, we observe that the test set in setting D, with inconsistent labels for images and texts, is more challenging than the test set in previous settings. As can be seen, most methods have an advantage over unimodal DenseNet and BERT. The Cross-Attention method provides better results for text, and Self-Attention method provides better results for images. The Self-Cross-Attention, on average, provides comparable results to the Self-Attention and Cross-Attention methods for both the modalities. Note that in all three attention methods, the multimodal-SSE technique has been used, which provides additional training data (with both consistent and inconsistent labels).

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[2] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 1

[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[4] Ilija Ilievski and Jiashi Feng. Multimodal learning and reasoning for visual question answering. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 551–562. Curran Associates, Inc., 2017. 1