# Multi-task Multimodal Learning for Disaster Situation Assessment

Tianyi Wang[1], Yudong Tao[2], Shu-Ching Chen[1], Mei-Ling Shyu[2]
[1]School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
[2]Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33124, USA
wtian002@cs.fiu.edu, yxt128@miami.edu, chens@cs.fiu.edu, shyu@miami.edu

***Abstract*—During disaster events, emergency response teams need to draw up the response plan at the earliest possible stage. Social media platforms contain rich information which could help to assess the current situation. In this paper, a novel multi-task multimodal deep learning framework with automatic loss weighting is proposed. Our framework is able to capture the correlation among different concepts and data modalities. The proposed automatic loss weighting method can prevent the tedious manual weight tuning process and improve the model performance. Extensive experiments on a large-scale multimodal disaster dataset from Twitter are conducted to identify post-disaster humanitarian category and infrastructure damage level. The results show that by learning the shared latent space of multiple tasks with loss weighting, our model can outperform all single tasks.**

***Keywords*-multi-task; multimodal learning; disaster information management;**

## I. Introduction

In recent years, due to the increase of human activities and the expansion of human habitat, the damages caused by natural disasters have been dramatically increased [1]. This poses a great challenge for disaster response management, which demands a more efficient and effective emergency response and recovery plan. Social media platforms have dramatically changed the way of sharing and acquiring information [2]. During the disastrous events, it can provide valuable situation updates which help the emergency response personnel to make faster and more accurate decisions. With the help of artificial intelligence (AI) techniques such as machine learning and deep learning, the emergency response team can analyze the social media data in real-time to determine the impact in different regions and prioritize the resource distribution accordingly. Many studies have been conducted to utilize social media data and deep learning methods to identify humanitarian activities, assess infrastructure damages, and enhance emergency awareness [3] [4] [5].

Regardless of the applications and methods used, lots of existing work on disaster situation assessment only focus on a single task and use a single data modality. One major limitation of such approaches is that the shared representations among different tasks and input modalities are not fully utilized. The main contributions of this paper include: (1) solving multiple tasks on a single model; (2) using the rich multimedia data from social media to train a learner with a better generalization ability on the problem; and (3) utilizing a novel loss weighting method for multi-task learning. The model can automatically adjust the loss weights for each task to improve the performance. The proposed method is more efficient and avoids the need for tedious manual weight tuning. We demonstrate the effectiveness of our proposed loss weighting method and compare its performance with models trained without loss weighting.

The remainder of the paper is structured as follows. Section 2 provides a brief literature review on disaster aituation assessment and multi-task learning. In Section 3, the proposed framework is presented. The experimental setup and results are discussed in Section 4. In Section 5, the conclusion and potential future work are given.

## II. Related work

### A. Disaster Situation Assessment Using Social Media Data

With the emerging use of social media, the researchers have access to timely information about disaster situations for humanitarian activity identification, infrastructure damage assessment, and emergency awareness enhancement. Li et al. [4] developed a convolutional neural network to categorize the damage level of the infrastructures in the images in three levels and detect the regions of damages based on the class activation map. In addition to image data, different types of data on the Internet can be informative to evaluate the disaster situation as well. In [6], a multimodal classification model was developed to classify disaster-related videos based on their contents in frames, audio, and descriptive texts. Since one video can include more than one type of contents, the analysis was further improved by treating the task as multi-label classification, which generates the in-depth analysis of the video data [7].

### B. Multi-Task Learning

Multi-task learning aims to train a single model for multiple related tasks with better learning efficiency and model performance than training one model for one task. The intuition behind multi-task learning is that the knowledge

learned from one task can be complementary for learning others. One main approach in multi-task learning is to integrate different task objectives and jointly train the model. [8] adopted this idea to train an encoder-decoder model for both distance estimation and building segmentation on remote sensing imagery. The performance metrics of both tasks are improved by utilizing multi-task learning.

## III. Proposed Framework

### A. Multimodal Learning

**Image Feature Extraction:** The VGG-19 [9] model pre-trained on ImageNet [10] is used to extract the low-level features from the imagery data. The pre-trained VGG-19 model serves as a feature extractor to generate the low-level features, which can be used by the rest of the layers in the model. We remove the last 2 fully connected layers in the original VGG-19 model and obtain the intermediate results from the last convolutional layer. During the training process, the weights of all the layers are kept fixed in the pre-trained model and fine-tuned on the rest of the fully connected layers.

**Text Embedding:** The proposed text embedding module is built upon the work of Peters *et al.* [11]. First, each character in the text corpus is tokenized and a temporal convolutional neural network (CNN) [12] is used to generate the CNN character embedding. The temporal convolutional module computes the 1-D convolution on the input data and generates the raw word vector of the input token. Second, the raw text vector is fed into two consecutive bi-directional long-short-term-machines (Bi-LSTMs). The forward and backward passes of the Bi-LSTM layer will learn the context information before and after the target word. In addition, a skip connection is added between the two Bi-LSTM layers. Finally, the weighted sum of the raw input vectors and the two intermediate vectors is calculated to form the final representation of the target word (as shown in Equation (1)).

$$V_k = \beta_k \cdot (s_0 \cdot x_k + s_1 \cdot h_k^1 + s_2 \cdot h_k^2), \tag{1}$$

where $V_k$ is the final word representation, $\beta_k$ is a task-specific scaling factor to help the model optimization, $s_i$ is the softmax-normalized weights, $x_k$ is the raw input vector for word $k$, and $h_k^1$ and $h_k^2$ are the intermediate vectors generated by the two Bi-LSTM layers.

**Multimodal Fusion:** In many tasks, data from different modalities could provide complementary information that helps generate a better generalization of the problem. Multimodal learning is utilized by fusing the early output from the text and image modules. The intuition of this approach is to preserve the semantic correlations from each data source. Features generated by the text and image modules are flattened and concatenated to form a joint vector.

### B. Multi-task Learning

Multi-task learning is adopted to improve the humanitarian activity and infrastructure damage classification tasks. The multi-task module implements the hard-parameter sharing methodology, in which all tasks share the same set of hidden layers. As a result, our model becomes more resilient to overfitting because the network is forced to learn multiple tasks simultaneously.

In multi-task learning, the loss function is defined in Equation (2).

$$L_{total}(x; W) = \sum_{i=1}^{N} w_i L_i, \tag{2}$$

where $L_{total}$ is the final loss that the model tries to optimize, $x$ and $W$ are the input and weight parameter to the model, $L_i$ and $w_i$ are the loss and corresponding weight scalar for task $i$, and $N$ is the total number of tasks. By default, the loss weights are uniformly assigned, which means all tasks are weighted at the same scale. However, this may cause the easier tasks to dominate the learning process. Also, when tasks use different loss functions, the loss for each task can have a significant difference due to how each function is mathematically defined.

In this paper, a novel loss weighting method is proposed that automatically adjusts the weighted scalar for each task. The goal is to adjust the loss of each task to a similar scale so that they can be optimized more equally during the training. The loss weights are determined based on the mean training loss. The rationale behind this is that a task with a higher total loss should also have a greater impact on the final weighted sum loss. By suppressing these tasks, the model can obtain a more balanced loss distribution at the beginning of the next training iteration. The details are illustrated in Algorithm 1. In steps 1 and 2, the iteration index starts at 0, and the validation loss $J^0$ and loss weights of each task $\hat{w}^0$ are initialized as infinite and all 1s, respectively. Then, in each iteration $t$, the training process $\mathrm{ModelTrain}()$ takes the loss weights from last epoch and generates the new validation loss $J^t$ and training losses $L_{i,j}^t$, where $i$ is the task index and $j$ is the epoch index. From steps 6 to 8, the mean training loss of each task $\overline{L}_i^t$ over all epochs is computed (for all $i = 1, \ldots, N$). Afterwards, the loss weights of each task $\hat{w}_i^t$ are computed by dividing the largest mean training loss by the mean training loss of the corresponding task. The iterative training will end when the validation loss stops decreasing.

## IV. Experiments and Analysis

### A. Dataset

To test the proposed framework, the CrisisMMD dataset [13] is used. This dataset contains 16,097 tweets and 18,126 images, and each tweet is associated with at least one image.

**Algorithm 1** The proposed loss weighting algorithm

**Input:** Training dataset $D_{train}$, validation dataset $D_{valid}$, and the number of training epochs $E$

$t \leftarrow 0$, $J^0 \leftarrow \infty$
$\hat{w}^0 \leftarrow [1, 1, ..., 1]$
**repeat**
  $t \leftarrow t + 1$
  $J^t, \{L^t_{i,j}\} \leftarrow \text{ModelTrain}(\hat{w}^{t-1}, D_{train}, D_{valid})$
  **for** $i = 1, \ldots, N$ **do**
    $\overline{L}^t_i \leftarrow \frac{\sum_{j=1}^{E} L^t_{i,j}}{E}$
  **end for**
  **for** $i = 1, \ldots, N$ **do**
    $\hat{w}^t_i \leftarrow \frac{\max_{k \in 1,\ldots,N} \overline{L}^t_k}{\overline{L}^t_i}$
  **end for**
  $\hat{w}^t \leftarrow [\hat{w}^t_1, \hat{w}^t_2, \ldots, \hat{w}^t_N]$
**until** $J^t > J^{t-1}$

Table I
PERFORMANCE COMPARISON BETWEEN SINGLE TASKS AND OUR METHOD. COLUMNS T (TEXT) AND I (IMAGE) INDICATE INVOLVED MODALITY. COLUMN O INDICATES THE APPLIED TASK, WHICH "INFO" IS INFORMATIVE CLASSIFICATION, "HUMAN" IS HUMANITARIAN CLASSIFICATION, AND "DAMAGE" IS DAMAGE CLASSIFICATION. COLUMN P INDICATES OUR PROPOSED MODEL.

| P | O | T | I | Precision | Recall | F1 | HL |
|---|---|---|---|---|---|---|---|
| | info | x | | 0.81 | 0.81 | 0.81 | 0.12 |
| x | info | x | | **0.83** | **0.83** | **0.83** | **0.11** |
| | info | | x | 0.66 | 0.62 | 0.64 | 0.22 |
| x | info | | x | **0.81** | **0.81** | **0.81** | **0.13** |
| | human | x | | 0.72 | 0.57 | 0.63 | 0.08 |
| | human | | x | 0.74 | 0.27 | 0.39 | 0.09 |
| | human | x | x | 0.72 | 0.57 | 0.63 | 0.11 |
| x | human | x | x | **0.73** | **0.58** | **0.64** | **0.10** |
| | damage | | x | 0.82 | 0.79 | 0.80 | 0.05 |
| | damage | x | x | 0.85 | 0.80 | 0.82 | 0.05 |
| x | damage | x | x | **0.88** | **0.82** | **0.86** | **0.05** |

Each sample is associated with multiple labels which contain its data informative indicator, humanitarian categories, and damage severity assessment. These labels contain the informativity of the text and image, type of humanitarian activities, and the level of damages appeared in the tweet. To utilize both text and image data for identifying the humanitarian category, a multi-label multimodal classification task is included to predict the combined humanitarian category labels from the two modalities. A similar multimodal approach is applied to the infrastructure damage level classification task, but we only did this for single label classification.

### B. Experimental Setup

The dataset is randomly split into 60% for training, 20% for validation, and 20% for testing. Categorical cross-entropy is used as the loss function for all single modality classification tasks, and binary cross-entropy is used for the multi-label classification task. Adam algorithm is chosen as the optimizer and the initial learning rate is set to 0.001. Each fully connected layer uses ReLu as the activation function with 50% dropout. Regarding the last fully connected layer, softmax is used for the single modality tasks, and sigmoid is used for the multimodal tasks. Precision, recall, F1 measure, and Hamming loss (HL) are used as the evaluation metrics. Our proposed multi-task multimodal model is trained on text informative classification, image informative classification, multimodal-multilabel humanitarian category classification, and multimodal infrastructure damage classification tasks.

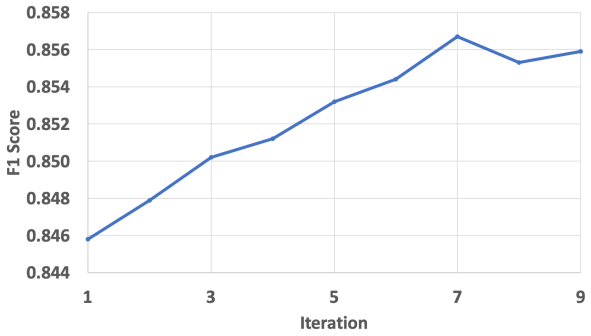


Figure 1. Multimodal damage classification results using the loss weighting algorithm

### C. Experimental Results

To demonstrate the effectiveness of our proposed model, it is compared with several baselines. Each baseline is a model trained on a specific task only. They include models trained on informative, humanitarian, and damage classification tasks using either text or image. Additionally, to compare the effectiveness of multimodal learning, two baseline models trained on humanitarian and damage tasks using both image and text are also added. Table I shows the performance comparison between the baselines and our proposed multi-task multimodal model.

The models are grouped based on the task and their performance is presented. In the first two groups of comparisons, our approach outperforms every baseline model on their corresponding tasks. These results illustrate the effectiveness of multi-task learning in the disaster situation assessment applications. In the next two groups of comparisons, all multimodal tasks achieved better performance when compared to their corresponding single modality tasks. This shows that multimodal learning is able to leverage the complementary information from the text and image inputs. Furthermore, the performance of the proposed loss weighting method is also evaluated. Figure 1 shows the F1 score of the multimodal damage level classification task at each iteration. After applying the loss weighting approach, the target task has

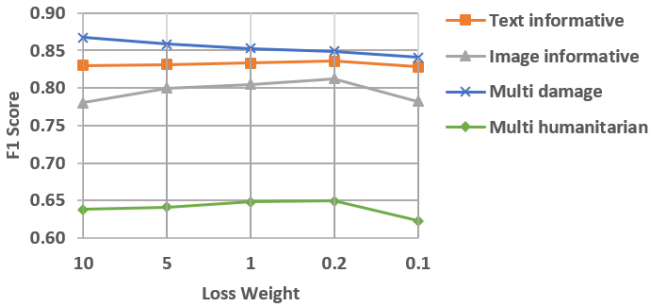


Figure 2. Performance comparison of all tasks when different loss weight settings for the multimodal damage level classification task.

shown a gradually improved F1 score from 0.846 to 0.857, which illustrates its effectiveness. Note that the scores started to oscillate after 7 iterations, which was mainly caused by the fact that the model became saturated.

To further investigate the effects of the loss weighting method, a 5-point sensitivity analysis test is conducted on the multimodal damage classification task. The goal of this test is to observe the task performance changes under different loss weight settings. The loss weight for the damage task is set to 10, 5, 1, 0.2, and 0.1 during the 5 runs. On the other hand, the loss weights of the other 3 tasks are set to 1 and did not change during the test. The results are shown in Figure 2. It can be observed that the performance of the damage task is consistently better with a higher loss weight, which is not surprising. In the contrary, the other 3 tasks get a slight performance boost when the loss weight of the damage task decreases. However, their performance starts to drop when the weight of the damage task becomes too small (0.1). Such performance boost can be explained by the smaller proportion of the damage task loss in the total loss, in which the model can generalize better on the entire problem domain. At the end, when the loss weight of the damage task is set to a very small value, the model can barely learn from this task, which causes the overall performance to be degraded.

## V. Conclusion and Future Work

In this paper, a novel deep learning framework is proposed based on multi-task and multimodal learning for social media disaster situation assessment. Furthermore, a loss weighting method is proposed to automatically adjust the loss weights for each task after every training cycle. The proposed model is evaluated on a multimedia natural disaster dataset collected from Twitter, and the experimental results showed that multi-task multimodal learning can improve the model performance by simultaneously learning the related tasks. Moreover, the proposed automatic loss weighting method is able to further improve the model performance. In future work, the loss weighting method can be improved by updating the loss weight after each training epoch, and our proposed method will be tested on non-disaster datasets.

## Acknowledgment

This research is partially supported by NSF OIA-1937019.

## References

[1] J. Klomp, "Economic development and natural disasters: A satellite data analysis," *Global Environmental Change*, vol. 36, pp. 67–88, 2016.

[2] S. Pouyanfar, Y. Yang, S. Chen, M. Shyu, and S. S. Iyengar, "Multimedia big data analytics: A survey," *ACM Comput. Surv.*, vol. 51, no. 1, pp. 10:1–10:34, 2018.

[3] S. Cresci, A. Cimino, F. DellOrletta, and M. Tesconi, "Crisis mapping during natural disasters via text analysis of social media messages," in *International Conference on Web Information Systems Engineering*, pp. 250–258, Springer, 2015.

[4] X. Li, D. Caragea, H. Zhang, and M. Imran, "Localizing and quantifying damage in social media images," in *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining*, pp. 194–201, 2018.

[5] J. Yin, S. Karimi, A. Lampert, M. A. Cameron, B. Robinson, and R. Power, "Using social media to enhance emergency situation awareness: Extended abstract," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 4234–4239, 2015.

[6] H. Tian, Y. Tao, S. Pouyanfar, S. Chen, and M. Shyu, "Multimodal deep representation learning for video classification," *World Wide Web*, vol. 22, no. 3, pp. 1325–1341, 2019.

[7] S. Pouyanfar, T. Wang, and S. Chen, "A multi-label multimodal deep learning framework for imbalanced data classification," in *2nd IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 199–204, 2019.

[8] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," *CoRR*, vol. abs/1709.05932, 2017.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations*, 2015.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[11] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 2227–2237, 2018.

[12] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, pp. 649–657, 2015.

[13] F. Alam, F. Ofli, and M. Imran, "Crisismmd: Multimodal twitter datasets from natural disasters," in *12th International Conference on Web and Social Media*, pp. 465–473, 2018.