Exploring Multimodal Features and Fusion Strategies for Analyzing Disaster Tweets

Raj Ratn Pranesh
Pennsylvania State University
rrp5338@psu.edu

Abstract

Social media platforms, such as Twitter, often provide firsthand news during the outbreak of a crisis. It is essential to process these facts quickly to plan response efforts in a manner that minimizes loss. In this paper, we present an analysis of various multimodal feature fusion techniques to analyze and classify disaster tweets into multiple crisis events via transfer learning. In our study, we utilized three image models pre-trained on ImageNet dataset and three fine-tuned language models to learn the visual and textual features of the data and combine them to make predictions. We have presented a systematic analysis of multiple intra-modal and cross-modal fusion strategies and their effect on the performance of the multimodal disaster classification system. In our experiment, we used 8,242 disaster tweets, each comprised of image and text data with five disaster event classes. The results show that the multimodal with transformer-attention mechanism and factorized bilinear pooling (FBP)(Zhang et al., 2019) for intra-modal and cross-modal feature fusion respectively achieved the best performance.

1 Introduction

The sudden breakout of crisis events, like natural disasters, creates high-stakes circumstances that are coupled with great uncertainty as well as the need to make quick decisions, often with limited official newscasts. Research in recent years has uncovered the importance of social media communication in disaster situations and shown that information broadcast via social media can improve situational awareness during an emergency (Vieweg et al., 2010). Social media has proven to be an active communication channel, especially during crisis events such as natural disasters including earthquakes, floods, and typhoons ((Hughes and Palen, 2009), (Imran et al., 2016)) or other emergencies such as accidents. These events spur a sudden surge of attention followed by reactive actions from both the general public and the media. The quick detection and analysis of such events are critical to swiftly disseminate information and, more importantly, prepare the relief team. Such situational awareness and tactical information enables the team effectively estimate early damage and launch relief efforts accordingly.

An automated system for crisis-related information retrieval from social media is imperative to rapidly and systematically classify disasters. Information regarding crises is best sourced from the social media site Twitter, which is a real-time, open, and public communication platform. The development of a system requires the extraction of relevant tweets to then classify them into different types of information: affected individuals, infrastructural damages, casualties, donations, caution, or advice. However, because the messages generated during a disaster vary greatly in value and since Twitter is a highly diverse platform, an automatic system needs to filter out messages that are irrelevant and do not contribute to situational awareness. As a result, we designed a system for detecting informative messages that classifies them to decide the type of information to extract (e.g., donation offers, casualty reports).

Information on social media mainly consists of textual messages and images. Past research has mainly focused on using textual content to aid disaster response. However, recent studies have revealed that images shared on social media during a disaster event can also help the relief team in several ways. For example, (Nguyen et al., 2017) incorporated images shared on Twitter to assess the severity of infrastructure damage in their work. Similarly, (Jing et al., 2016) investigated the usefulness of images and text for their study on flood and flood aid. Our work follows this method of taking into account both texts and images.

Previous works (Ofli et al., 2020), (Agarwal

Proceedings of the 2022 COLING Workshop: The 8th Workshop on Noisy User-generated Text (W-NUT 2022), pages 62–68

et al., 2020), (Kumar et al., 2020), (Abavisani et al., 2020) have proposed a multimodal system for analyzing disaster tweets that utilizes feature fusion. However, not much exploration has been done for the enhancement of the extracted visual, textual and their combined multimodal feature representation. In this paper, we present an analysis of various multimodal fusion strategies for intra-modal fusion and cross-modal fusion. We investigate relation-attention, self-attention, and transformer-attention for intra-modal fusion. For the cross-modal fusion, we explore three methods, namely, Factorized Bilinear Pooling (Zhang et al., 2019), Compact Bilinear Gated Pooling (Kiela et al., 2018) and Compact Bilinear Pooling (Fukui et al., 2016). Along with this, we evaluate state-of-the-art models which were three pretrained image models (VGG19 (Simonyan and Zisserman, 2014), ResNet-50 (He et al., 2016) and AlexNet) and three pretrained language models (BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019)) for the disaster tweet analysis and classification task. In our analysis, we utilize multimodal CrisisMMD (Alam et al., 2018) dataset. We found that the ResNet-50 outperformed other image models and among the textual models RoBERTa achieved the best performance. We further utilize these two models for the evaluation of intra-modal and cross-modal fusion strategies.

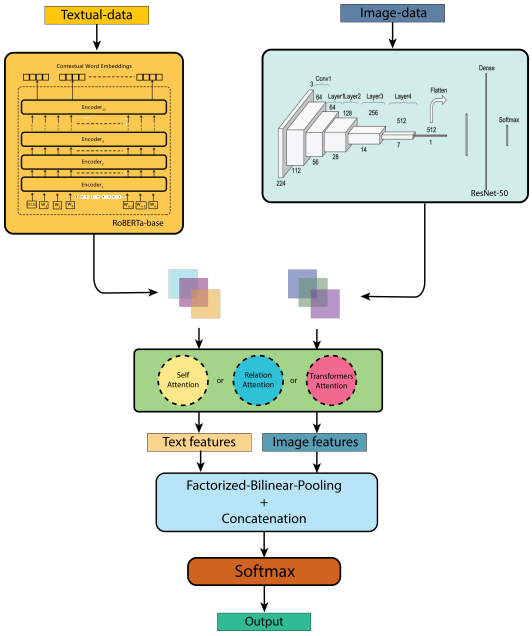

Figure 1: Feature fusion pipeline with textual sub-model (RoBERTa) and visual sub-model (ResNet-50)

2 Methodology

2.0.1 Textual feature extractor:

We employed three pretrained language models, namely, BERT-base (Devlin et al., 2018), RoBERTa-base (Liu et al., 2019) and ALBERT-base (Lan et al., 2019) to extract a high quality text feature vector. We finetuned them with a custom classification head with updatable weights. The averaged pool of sequential output from 12 encoding layers of each model was used as the custom classifier head's input. Once the model was fine-tuned, each of the language models was fed with a sequence of text inputs (reprocessed tweets) which then went through all of the stacked encoding layers, thereby extracting essential features from the context.

2.0.2 Visual Feature Extractor:

For image feature extraction, we use three image models, namely, VGG19 (Simonyan and Zisserman, 2014), ResNet-50 and AlexNet pretrained on Imagenet-21k (Deng et al., 2009). Each of the pre-trained image models was supplied with a pre-processed image; a visual representation was then extracted from the final finetuned FC layer of each model. The output is a vector of the dimension of 4096, 1000, 1000 for VGG19, ResNet-50 and AlexNet respectively.

2.1 Multimodal fusion

2.1.1 Intra-modal feature fusion

We have developed functions using different attention-based methods, namely, self-attention, relation-attention and transformer-attention methods. These functions can convert a variable number of features into a fixed dimension feature. For an "n" number of features, we denote the i_{th} feature as f_i where $i \in [1, n]$. We applied fusion techniques as follows:

- $Self - attention$: For each feature we apply a 1-dimensional fully connected layer $W^0_{d\times 1}$ and a sigmoid function σ, resulting to the weight a_i of the i_th feature f_i^T as follows:

$$\alpha_i = \sigma(f_i^T \cdot W^0_{d\times 1}) \quad (1)$$

We combined these weights from self-attention (Vaswani et al., 2017) for every feature into a global representation f_s as follows:

$$f_s = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i} \quad (2)$$

- $Relation - attention$: The function derives the relationship between the features and generates relevant features. Since f_s holds global representation of these features, we use sample concatenation of each feature and global representation to shape the global-local relation[f_i:f_s]. Next, we apply the 1-dimensional fully connected layers $W^1_{d/times1}$ with the sigmoid function σ. For relation-attention weight β of i_th feature $[f_i : f_s]^T$ is computed as:

$$\beta_i = \sigma([f_i : f_s]^T \cdot W^1_{d/times1}) \quad (3)$$

Using aggregated weights from self-attention function and relation-attention function, we combine all the features to get a new feature f_r:

$$f_r = \frac{\sum_{i=1}^{n} \alpha_i \beta_i [f_i : f_s]}{\sum_{i=1}^{n} \alpha_i \beta_i} \quad (4)$$

- $Transformer - attention$: Based on the works in (Zhang et al., 2019) and (Yang et al., 2016), we compute the attention weight as follows:

$$f'_i = W^2_{m \times d} \cdot f_i + b\gamma_i \quad (5)$$

$$= exp(u^t_{d \times 1} \cdot tanh(f'_i)) \quad (6)$$

To reshape the the dimension of feature f_i, we feed it through a $w \times d$ dimensional FC layer 6. The weight of i_{th} feature f_i is processed through the tanh function which is then fed to the exp function along with dot product of $u^t_{d \times 1}$. We pass the output from the exp function to a 1-dimensional FC layer stated in 6. From the transformers attention we formulate all the features into a single feature f_i, as

$$f_s = \frac{\sum_{i=1}^{n} \gamma_i f_i}{\sum_{i=1}^{n} \gamma_i} \quad (7)$$

2.1.2 Cross-modal feature fusion

- $FactorizedBilinearPooling(FBP)$ (Zhang et al., 2019): The two feature vectors obtained via different modalities are fused together by applying FBP function.

- $CompactBilinearPooling(CBP)$: Originally proposed (Fukui et al., 2016) for VQA task, we modified this feature fusion technique for the classification task.

- $CompactBilinearGatedPooling(CBGP)$: With an additional attention gate applied on top of the compact bilinear pooling module, we adopted the CBGP (Kiela et al., 2018) fusion technique for the cross-modal feature fusion.

3 Dataset

We have used the CrisisMMD (Alam et al., 2018) dataset for training and testing our model. Each text and image pair in the dataset have two annotations: (task_1) humanitarian categories (eight classes), (task_2) informative vs. not-informative (two classes). Since the number of labels across different classes was uneven, following (Ofli et al., 2020), we compressed the number of humanitarian categories to five- namely, (i) *Not-humanitarian* (4312), (ii) *other_relevant_information* (1764), (iii) *rescue_volunteering_or_donation_effort* (1195), (iv) *infrastructure_and_utility_damage* (842) and (v) *affected_individuals* (129). In the CrisisMMD dataset, tweet text and image in a pair were annotated separately, as a result, few pairs had a different label for text and it's associated image. We removed those pairs and performed the experiment only those data who have the same label for text and image. Finally, we have 8,242 pairs and split the data in 70%:15%:15% ratio for training (5770), development (1236), and test (1236) sets. For the informative and not-informative, we had 7875 (train), 1687 (development) and 1688 (test).

4 Experiment

4.1 Exploring Visual feature

In the visual modal, we compared three image models, namely: AlexNet (Krizhevsky et al., 2012),ResNet-50 (He et al., 2016) and VGG19 (Simonyan and Zisserman, 2014); pretrained on large ImageNet (Deng et al., 2009) dataset. In the visual unimodal for each of the image model,

the extracted feature vector was passed through two consecutive fully connected layers of dimension 512 and 256. The feature vector was then passed into a batch normalization layer and dropout layer(with dropout probability = 0.4), followed by a 5-dimensional dense layer with a softmax activation function in order to make the final class prediction of the disaster event. Relu activation function and L2 regularization of 0.01 was applied at each dense layer. All of the image models were trained on the training dataset(learning rate = 1e-4) using Adam (Kingma and Ba, 2014) optimizer and with cross-entropy as the loss function. The model's hyperparameter fine-tuning was done on the validation set. We also conducted an evaluation of three models over the test dataset. As shown in table 1, out of all three image models, ResNet-50 achieved the best F1 score of 68.35 as compared to ResNet-50 (He et al., 2016) and AlexNet. This shows that the ResNet-50 was able to understand the image feature more clearly and generate better image representation. The reason behind this could be the residual module based ResNet-50's deeper architecture which lacks in VGG19 and AlexNet models.

Model	Precision	Recall	F1-score
AlexNet	74.42	56.74	64.38
VGG19	76.39	55.01	63.96
ResNet-50	79.23	60.11	68.35

Table 1: Performance of image unimodal on task_1

4.2 Exploring Textual feature

Similar to the visual modal, the textual modal utilizes transfer-learning for learning the textual data representation. For the textual unimodal, we applied the bidirectional transformers with the self-attention mechanism to extract resourceful features from text in the disaster tweets. In our analysis, we use ALBERT-base (Lan et al., 2019), BERT-base (Devlin et al., 2018) and RoBERTa-base (Liu et al., 2019) pretrained language models. These models are mainly known for their pretrained weights over different domain data. For our task, we fine-tuned all of the models on the disaster dataset. As we discussed above, the input text sequence was structured, tokenized and pre-processed according to the language model's input format. From each of the language models, we extracted the $[CLS]$ (for BERT and ALBERT) or $< s >$ (for RoBERTa) which represents the entire input sentence and is used as the aggregate sequence representation for classification tasks. Similar to the visual unimodal, the classification token was then passed through a series of the fully connected layer of size 512 and 256. This was followed by a batch normalization layer, dropout layer(dropout probability = 0.4), and a 5-dimensional dense layer with a softmax activation function. All the dense layer in the model has a relu activation function and L2 regularization of 0.01. All of the models were trained with the learning rate of 1e-4, using Adam (Kingma and Ba, 2014) as optimizer and cross-entropy as the loss function. On analyzing the performance of all the three models on the test data, we observed (table 2) that the performance of RoBERTa-base unimodal was the most optimal. BERT and ALBERT achieved the F1 score of 72.92 and 71.23 respectively.

Model	Precision	Recall	F1-score
ALBERT-base	77.34	66.02	71.23
BERT-base	79.34	67.47	72.92
RoBERTa-base	85.36	66.2	74.56

Table 2: Performance of Text Unimodal on task_1

4.3 Exploring Fusion Strategies

Feature extraction: We extracted the feature maps from the preprocessed visual and textual data and utilized them for the intra-modal fusion. For a given 3 dimension feature map, the size is represented as $H \times W \times C$, where H and W represented the height and width of the feature map, respectively. The number of channel in the feature map was represented as C. For the intra-modal fusion process, we sliced the feature map into n vectors such that $n = H \times W$. Therefore, n number of C-dimensional vectors were obtained. For the image data, we extracted the feature map from the layer before the final average polling layer of the ResNet-50. For the RoBERTa model, instead of using classification token, we extracted the vector sequence consisting of each input token's vector representation. The size of each output token sequence was 768 x 42 (max_length). This vector was split into 768 feature vector (42-dimensional) before intra-modal fusion.

Intra-modal Fusion: As we discussed above in the section *Multimodal Fusion*, we utilized 3 intra-modal attention fusion methods: relation-attention,

self-attention, and transformer-attention. Both the visual and textual feature vector were subjected to each of the attention methods before performing the cross-modal fusion. The *n* split feature vectors from each of the visual and textual modalities, when passes through the attention layer, condenses to form respective unique representations which are then use for the cross-modal fusion.

Cross-modal Fusion: For the cross-modal fusion, we investigated 3 methods: factorized bilinear pooling, compact bilinear pooling and compact bilinear gated pooling. The visual and textual feature vector generated after the intra-modal fusion is then subjected to cross-modal fusion to produce a combined multimodal representation. The multimodal vector is then passed through a classification layer of size 5 with a softmax activation function to make predictions. The model is trained on a batch size of 64 with cross-entropy loss function and Adam (Kingma and Ba, 2014) optimizer for training the model. During the training of the model, we use an initial learning rate of 1e-5, two callback API-early-stopping conditions and reduce the learning rate on the plateau (reducing factor = 0.5, patience = 5).

Textual \ Visual	**Self attention**	**Relation attention**	**Transformers attention**
Self-attention	78.7%	79.4%	81.7%
Relation-attention	79.9%	81.1%	82.2%
Transformers-attention	80.0%	81.2%	85.1%

Table 3: Multimodal performance (macro F1 %) on task_1 with FBP

Textual \ Visual	**Self attention**	**Relation attention**	**Transformers attention**
Self-attention	82.8%	83.1%	84.9%
Relation-attention	81.8%	84.3%	85.1%
Transformers-attention	82.1%	85.2%	89.5%

Table 4: Multimodal performance (macro F1 %) on task_2 with FBP

5 Results

In this section, we discuss and analyze the multimodal performance with various fusion techniques. Table 3 and 4 show the Macro F1-score of FBP fusion methods on task_1 and task_2 respectively. We have shown the result of the **best cross-model fusion method**: **FBP** applied with various intra-model fusion methods.

For task_2, we observed that by using the FBP (Zhang et al., 2019) and Transformer attention layer in the pipeline, the performance of multimodal was remarkably better (around **12%**) than the other cross layer fusion methods (CBP and CBGP). We also noticed that in either of the cross-modal fusion method, the transformer attention intra-modal fusion performed the best. For task_1 (refer 3) and task_2 (refer 4), FBP with transformers-attention based multimodal model gave the best result of 85.1% and 89.5% respectively. We can also see that models having transformer-attention combined with relation-attention outperformed the model with transformer-attention and self-attention.

Coming to the multimodal baseline (Ofli et al., 2020) and (Abavisani et al., 2020), our model outperform it by **7.99%** and **1.10%** on the task_1 and for task_2 it is **5.92%** and **0.78%**. The reason behind the superior performance of our model lies behind the underlying feature representation generated by the pre-trained language and image models. Moreover, we were able to capture intra-modality information using attention mechanism which produced a denser feature representation before the cross-modal fusion. Therefore using transfer learning and attention-based fusion techniques, we were able to blend together with powerful language and image models and build a more robust multimodal.

6 Conclusion

In this paper, we present an extensive analysis of multiple feature fusion strategies for developing a multi-modal framework for detecting and classifying tweets into various crisis events accurately based on the textual and visual features. In our study, we compared various image and language models and found that the ResNet and RoBERTa outperformed the other models. We also presented a comparative study of various fusion methods; through that, we can conclude that the selection of effective intra-modal and cross-modal method plays a crucial role in developing a more accurate and efficient multimodal framework for classifying the events for faster relief efforts. We observed that the transformer-attention mechanism outperformed the other intra-modal fusion methods. We also showed that by using factorized bilinear pooling, the multimodal feature representation can be improved. The results of the experiments show that one application of the multimodal framework can be the identification and filtration of disaster-related information available on social media platforms.

References

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

Mansi Agarwal, Maitree Leekha, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 346–353.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth International AAAI Conference on Web and Social Media*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

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

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Amanda Lee Hughes and Leysia Palen. 2009. Twitter adoption and use in mass convergence and emergency events. *International journal of emergency management*, 6(3-4):248–260.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.

Min Jing, Bryan W Scotney, Sonya A Coleman, Martin T McGinnity, Xiubo Zhang, Stephen Kelly, Khurshid Ahmad, Antje Schlaf, Sabine Gründer-Fahrer, and Gerhard Heyer. 2016. Integration of text and image analysis for flood event image recognition. In *2016 27th Irish Signals and Systems Conference (ISSC)*, pages 1–6. IEEE.

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2018. Efficient large-scale multimodal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Abhinav Kumar, Jyoti Prakash Singh, Yogesh K Dwivedi, and Nripendra P Rana. 2020. A deep multimodal neural network for informative twitter content classification during emergencies. *Annals of Operations Research*, pages 1–32.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dat T Nguyen, Ferda Ofli, Muhammad Imran, and Prasenjit Mitra. 2017. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 569–576.

Ferda Ofli, Firoj Alam, and Muhammad Imran. 2020. Analysis of social media data using multimodal deep learning for disaster response. *arXiv preprint arXiv:2004.11838*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1079–1088.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Yuanyuan Zhang, Zi-Rui Wang, and Jun Du. 2019. Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.