Multimodal Learning for Disaster Assessment Using Social Media Data

Swetha Revanur and Keanu Spies (srevanur, keanus) @ stanford.edu

Introduction

- 2.5 quintillion bytes of data are created each day, largely in the form of images and text on social media platforms.

- Social media is particularly useful in disaster response and communication, and can be leveraged as a tool for real-time alerts in the case of natural disasters or wars.
- However, relief agencies and law enforcement officials struggle to assess their confidence in crowdsourced media, given the susceptibility of the platforms to spammers and adversarial attacks.

Objectives

Dataset

- We worked with the "Multimodal Damage Identification for Humanitarian Computing" dataset consisting of 5831 captioned images from Twitter, Instagram, and Google Images.
- Each image and caption has one of six labels: fires, floods, natural landscape, infrastructural, human, and non-damage.

Project Milestones

- <u>Phase 1</u>: Implement text-only and image-only baselines with both traditional approaches and deep neural networks to classify disaster types.
- <u>Phase 2</u>: Propose, design, and develop multimodal approaches for classification, spanning feature and decision fusion.

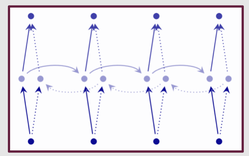
Phase 1: Unimodal Approaches

A. Text and Image Preprocessing

Each modality requires its own preprocessing steps. For text, we performed casefolding to convert all characters to lowercase. We removed all punctuation including "#" signs. Emojis and characters in other languages were retained throughout all steps. Images require different transformations. Given the range of resolutions, images were first resized to be 224 × 224, then normalized.

B. Text-Only Classifiers

- We explored multinomial naive Bayes, Bernoulli naive Bayes, linear kernel SVM, and extra-tree models with Word2Vec embeddings. We also considered term frequency-inverse document frequency (TF-IDF) variants of the above algorithms to take into account word importance.
- Furthermore, we trained word embeddings from scratch and implemented bidirectional LSTM and GRU networks with self-attention.



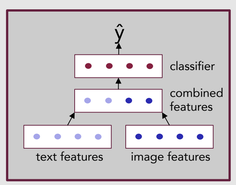
C. Image-Only Classifiers

- Here, we used transfer learning approaches for classification. We finetuned pretrained ResNet-18 and ResNet-50 models.

Phase 2: Multimodal Approaches

A. Feature Fusion

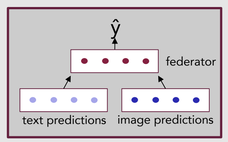
- Feature fusion combines the deep features from the bidirectional RNN and the ResNet-50 to train a multimodal classifier.



- The intermediate features are extracted from the second-to-last layer of each deep network. They are concatenated and used to train a one-layer classifier.

B. Decision Fusion

- Decision fusion combines the predictions of the top-performing unimodal classifiers (SVM with TF-IDF and ResNet-50) to obtain improved predictions with a federator.



- The federator is trained to determine whether, for a given complete social media post, an image-based or text-based classification is more likely to be correct.
- Specifically, the federator is passed an (image, text) tuple and is trained on a modified set of binary labels, where 1 indicates that the image-only classifier strictly outperforms the text-only classifier, and 0 is the catch-all for any other cases.

A Note on Training Models

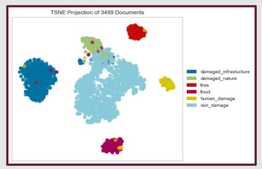
- A grid search was used to tune hyperparameters (i.e., initial learning rate, learning rate decay, dropout rate, batch size, and weight initializations) for all networks. Early stopping, batch normalization, and regularization were included to reduce variance.
- We used SGD with momentum to optimize the objective function.

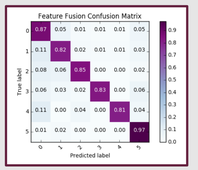
Results and Conclusions

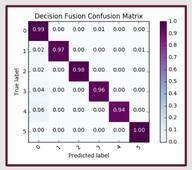
- The results on the test set are as follows:

Type	Model	Precision	Recall	F1-Score	Accuracy
Text	MultiNB	0.8769	0.6940	0.7748	0.8472
Text	MultiNB_TF	0.8882	0.4280	0.5777	0.7098
Text	BernNB	0.4786	0.2845	0.3569	0.6680
Text	BernNB_TF	0.4786	0.2845	0.3569	0.6680
Text	SVM	0.8565	0.8208	0.8383	0.8906
Text	**SVM_TF**	**0.8942**	**0.8256**	**0.8585**	**0.9052**
Text	Tree	0.5755	0.4687	0.5167	0.6853
Text	Tree_TF	0.6037	0.4872	0.5392	0.6906
Text	LSTM	0.8415	0.7144	0.7516	0.7144
Text	GRU	0.6933	0.6124	0.5231	0.6124
Image	ResNet-18	0.8196	0.8165	0.8161	0.8165
Image	**ResNet-50**	**0.8352**	**0.8345**	**0.8329**	**0.8345**
Multi	Feature Fusion	0.9158	0.9151	0.9146	0.9151
Multi	**Decision Fusion**	**0.9890**	**0.9889**	**0.9888**	**0.9889**

- Compared to the best unimodal model, decision fusion sees an 8% increase in classification accuracy. This demonstrates the utility of multimodal approaches.







Future Work

- We hope to explore online policy adaption algorithms for ensembling classifiers.
- In addition, to improve model explainability and as an alternative fusion method, we can implement coattention with the captions and images.

References

Hussein Mouzannar et al.. Damage Identification in Social Media Posts Using Multimodal Deep Learning, 2018.

Tom Zahavy et al.. Is a picture worth a thousand words? A Deep Multi-Modal Fusion Architecture for Product Classification in e-Commerce, 2018.