Multi-task Learning to Enable Location Mention Identification in the Early Hours of a Crisis Event

Sarthak Khanal
Kansas State University
sarthakk@ksu.edu

Doina Caragea
Kansas State University
dcaragea@ksu.edu

Abstract

Training a robust and reliable deep learning model requires a large amount of data. In the crisis domain, building deep learning models to identify actionable information from the huge influx of data posted by eyewitnesses of crisis events on social media, in a time-critical manner, is central for fast response and relief operations. However, building a large, annotated dataset to train deep learning models is not always feasible in a crisis situation. In this paper, we investigate a multi-task learning approach to concurrently leverage available annotated data for several related tasks from the crisis domain to improve the performance on a main task with limited annotated data. Specifically, we focus on using multi-task learning to improve the performance on the task of identifying location mentions in crisis tweets.

1 Introduction

Social media has evolved into a platform for people to share their concerns, report information as eyewitnesses of events, and also call for help, especially during crisis situations. The huge amount of data that is posted on social media during crisis events could be used to build reliable and robust deep learning models for identifying information useful to crisis management and response teams. However, using social media data for a particular task, oftentimes, requires intensive manual effort in the form of annotation. The effort becomes even more arduous when we consider the noisy nature of social media content and the amount of labelled data required for a typical deep learning model.

The domain of crisis-related social media analysis, tweets in particular, is a well-researched field with labelled data available for various tasks (Imran et al., 2016; Middleton et al., 2014; Alam et al., 2018). However, most of the available human-annotated datasets consists of thousands of instances, at best, which means that crisis datasets are relatively small compared to those available for tasks in other domains. Furthermore, for tasks that can support a new, emergent crisis situation, human-labelled data of large volume cannot be acquired for the reasons discussed above. In this work, we explore ways in which we can harness the available small datasets from the domain to bolster performance for individual tasks of interest.

One popular approach in addressing the size-limitation of labelled data for a particular task is to leverage unlabelled data. In the field of Natural Language Processing (NLP), the recent advancements in transformer-based architectures, and associated pre-training with huge amounts of unlabelled data, has largely been successful in addressing this issue. Transformer-based architectures, currently, hold state-of-the-art results for many NLP tasks. However, the domain shift from the pre-training corpus to a downstream task's domain is still a significant issue (Han and Eisenstein, 2019). Moreover, further pre-training with domain-specific unlabelled data is compute-intensive (Devlin et al., 2018) and it is not always feasible.

An alternative approach to address the limitation in terms of labeled data is to concurrently leverage smaller datasets available for different, but related tasks using a multi-task learning strategy (Caruana, 1997). In the multi-task setting, some layers can be shared across different tasks, while each task can also have one or more task-specific layers, and the entire model is trained in parallel for all the tasks. A multi-task model is designed with the intuition that the lower layers of the model learn abstract features common to related tasks, while the upper layers learn features specific to each individual task. This approach is especially useful in the domain of crisis-related social media analysis, given the lack of large datasets, while smaller datasets are available for different tasks.

In this work, our main focus is on the task of identifying fine-grained locations from tweet texts.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4051–4056
November 7–11, 2021.

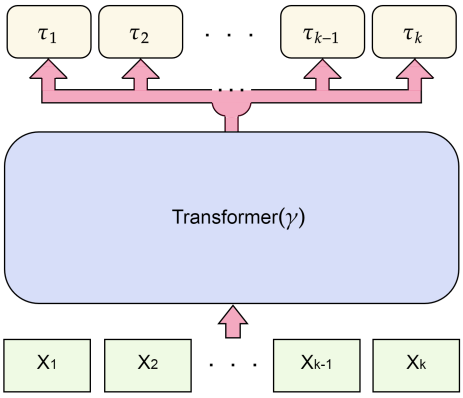

Figure 1: Multi-task model overview

Identification of location entities in tweets posted during crisis events is vital in extracting actionable situational awareness information. Furthermore, identifying the entities according to a hierarchy of location types can help in geographical location disambiguation and geo-coding. We use the English subset of the dataset published by Middleton et al. (2014), which has only a few thousands instances.

To address the issue of limited data size, we use a multi-task learning setting to augment the learning of fine-grained location identification with three other tasks in the domain of crisis-related tweets: key-phrase identification (Chowdhury et al., 2020), eyewitness-account classification (Zahra et al., 2020) and humanitarian categories classification (Alam et al., 2018). We hypothesize that the similar nature of the tasks and the common abstract objective of identifying actionable information in crisis-related tweets will result in a performance boost for the main task considered, i.e., identification of fine-grain locations in tweet texts.

2 Related Work

Multi-task learning is a well-researched topic in the field of NLP, and deep learning, in general (Ruder, 2017). Caruana (1997) outlines one of the popular strategies for implementing multi-task learning: hard parameter sharing. In hard parameter sharing, a module shared by all tasks is followed by task-specific modules. On the other hand, in soft parameter sharing (Duong et al., 2015; Yang and Hospedales, 2017), each task has its own set of layers, but the parameters of the task-specific layers are constrained to be similar across tasks to enforce exchange of information among tasks. Hard parameter sharing is known to reduce over-fitting (Ruder, 2017), and is thus useful for tasks with small training sets. The usefulness of multi-task learning has been shown in a variety of applications, including image (Zhang et al., 2014; Cheng et al., 2011), voice (Stoller et al., 2018; Rao et al., 2018) and text (Wang et al., 2020; Liu et al., 2019a; Pham et al., 2019) analysis. In the crisis domain, Wang et al. (2020) presented a multi-modal multi-task model using a single multi-modal dataset containing labels for different tasks. Chowdhury et al. (2020) used single-token keywords identification as an auxiliary task when predicting multi-token keyphrases. The work by Liu et al. (2019a) is the closest to our work, in that they used separate task-specific datasets in a multi-task learning setting. However, they performed experiments on general NLP benchmark datasets, such as GLUE (Wang et al., 2018).

3 Background and Approaches

We use the hard-parameter sharing approach for multi-task learning, where task-specific modules (τ) are attached on top of a shared module (γ) as shown in Figure 1. We use the algorithm proposed by Liu et al. (2019a) to train multiple task modules, in parallel.

Multi-task Learning. Multi-task learning can be formulated as follows: Given a set of k tasks and their corresponding data $D = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_k, Y_k)\}$, where $(X_i, Y_i) \in D$ is the training dataset for the i^{th} task, the goal of multi-task learning is to minimize the aggregate loss $L(\theta)$ on D given by:

$$L(\theta) = \sum_{i=1}^{k} \sum_{(x,y)\in(X_i,Y_i)} \ell^i(y, T^i(x, \theta)) \quad (1)$$

where ℓ^i is the loss function for the i^{th} task, $T^i(x, \theta) = \tau^i(\gamma(x, \theta_\gamma), \theta_i)$ is the output of the i^{th} task module τ^i (with parameter $\theta_i \subset \theta$) and γ is the shared module (with parameter $\theta_\gamma \subset \theta$).

Token Classification Tasks. A token classification task maps a sequence $x = \{x_1, \cdots, x_n\}$, representing n input tokens, to a sequence $y = \{y_1, \cdots, y_n\}$, representing n task-specific tags corresponding to the tokens in the sequence x. The mapping can be generically written as: $\tau^i(\theta_i) : \mathbf{x} \rightarrow \mathbf{y}$ (where θ_i are the task parameters). The loss ℓ^i that we use for these tasks is the negative log likelihood. In our setting, the fine-grained location identification and the keyphrase identification tasks are token classification tasks.

Sequence Classification Tasks. A sequence classification task maps a sequence of n tokens $x = \{x_1, \cdots, x_n\}$ to a class $y \in C$, where C is the task-specific set of classes. The mapping can be generically written as $\tau^j(\theta_j) : \mathbf{x} \rightarrow \mathbf{y}$ (where θ_j are the task parameters). We use the standard cross-entropy or the binary cross-entropy loss ℓ^j for sequence classification tasks, depending upon the number of classes. In our setting, the eyewitness-account classification and the humanitarian categories classification tasks are seen as sequence classification tasks.

Shared Module. For the shared module, we experiment with three base variants (hidden layer size of 768) of transformer-based language models: BERT (Devlin et al., 2018), Albert (Lan et al., 2019) and RoBERTa (Liu et al., 2019b). BERT is one of the first and most popular transformer-based models, and has been used widely in various applications. We compare its performance with that of RoBERTa, which was trained on a larger dataset than BERT and it is considered to be more robust (Liu et al., 2019b). We also compare BERT and RoBERTa with Albert, a significantly smaller model, to understand the effect that the size of the shared module has on the the performance of the task at hand.

4 Experimental Setup

4.1 Datasets

Fine-grained Location Identification. We use the dataset published by Middleton et al. (2014), which has two sets of tweets posted during Hurricane Sandy 2012 and Christchurch Earthquake 2012, respectively. Based on the place of the event, in what follows, the set of tweets posted during Hurricane Sandy 2012 will be referred to as *NY*, while those posted during Christchurch Earthquake 2012 will be referred to as *NZ*. The NY and NZ sets contain 1907 and 1762 unique tweets, respectively. Both the NY and NZ tweets have human-labelled location entities corresponding to three categories: *administrative location, building* and *transportation*. We randomly select 1000 tweets from each set for training, 500 tweets for test, and the remaining tweets for development (dev) splits.

Keyphrase Identification. We use the dataset published by Chowdhury et al. (2020) for keyphrase identification. The dataset contains tweets from various crisis events. The tokens in each tweet are labelled as keyphrase or not, using the script provided by Chowdhury et al. (2020). We use a random sample of 1000 tweets for training to keep the dataset size balanced across multiple tasks.

Eyewitness-account Classification. The dataset published by Zahra et al. (2020) contains tweets from flood, hurricane and earthquake. The tweets are labelled using one of five eyewitness classes: *direct-eyewitness, indirect-eyewitness, vulnerable direct-eyewitness, non-eyewitness*, and *don't know.* As for the other tasks, we use a class-balanced sample of 1000 tweets from the dataset for training.

Humanitarian Categories Classification. We use the dataset published by Alam et al. (2018) for humanitarian categories classification. The dataset contains tweets from various crisis events, labelled using the following humanitarian classes: *Infrastructure and utility damage, Vehicle damage, Rescue, volunteering, or donation effort, Injured or dead people, Affected individuals, Missing or found people, Other relevant information*, and *Not relevant or can't judge.* We use a class-balanced sample of 1000 tweets from the dataset for training.

4.2 Baseline and Metrics

We use a single task-setting, where only one task-module is attached over the base module, as the baseline to which we compare the performance of the multi-task learning setting. We use precision (Pr), recall (Re) and F1 scores as metrics to compare the performance of the models.

4.3 Experiments and Implementation Details

We perform the experiments on three levels of task-sharing: a **Single-task model (ST)** (trained on *NY* and *NZ*, respectively, and tuned/tested on the corresponding dev/test sets), a **Location-only Multi-task model (LMT)** (trained with *NY* and *NZ* as two different location identification tasks in a multi-task setting, and tuned/tested on the corresponding dev/test sets), and **Multi-task model (MT)** (trained with all tasks in a multi-task setting, and tuned/tested on the corresponding dev/test sets for the location identification tasks). In addition, we also combine the two location datasets *NY* and *NZ* (*Combined*) and train a single-task model (*ST*), and a multi-task model (*MT*), using the other three tasks to understand the benefit of only using the related tasks for general location identification (as opposed to disaster-specific identication). The hyperparameter configuration used in the experiments are shown in the Appendix A. All experiments are run on a 4-GPU system of NVIDIA Tesla V100 GPUs.

Task	γ	Dev-F1	Test		
			Pr	Re	F1
Dataset: NY					
ST	BERT	75.40	79.18	79.41	79.30
	RoBERTa	66.31	70.82	73.53	72.15
	Albert	72.40	78.72	76.18	77.43
LMT	BERT	75.49	76.09	**82.35**	79.10
	RoBERTa	69.69	81.72	69.71	75.24
	Albert	73.29	76.68	77.35	77.01
MT	BERT	73.95	78.90	80.29	**79.59**
	RoBERTa	70.12	76.95	69.71	73.15
	Albert	77.43	**84.00**	74.12	78.75
Dataset: NZ					
ST	BERT	67.68	58.55	66.86	62.43
	RoBERTa	71.25	66.67	57.99	62.03
	Albert	69.52	61.58	69.23	65.18
LMT	Bert	65.99	56.72	67.46	61.62
	RoBERTa	69.01	66.67	63.91	65.26
	Albert	70.83	70.35	**71.60**	**70.97**
MT	BERT	72.32	69.05	68.64	68.84
	RoBERTa	70.37	**72.79**	58.58	64.92
	Albert	75.14	67.47	66.27	66.87
Dataset: Combined					
ST	BERT	75.95	73.14	**85.07**	78.66
	RoBERTa	69.90	72.59	68.17	70.31
	Albert	68.92	71.72	68.76	70.21
MT	BERT	76.33	76.01	80.94	78.40
	RoBERTa	69.01	71.98	70.14	71.04
	Albert	78.65	**83.54**	77.80	**80.57**

Table 1: Results for Fine-grained location identification task in three settings: Single-task (ST), Location-only Multi-task (LMT) and all-multi-task (MT) with three variants of transformers (γ) as shared module. The results are grouped by dataset: NY, NZ and Combined.

5 Results and Discussion

Table 1 shows the results for the location identification task grouped by dataset (*NY*, *NZ* and *Combined*) in three different task-sharing settings, with three variants of the transformer-based shared module, as described in Section 4.3. The table shows precision, recall and F1 scores for the test set, along with F1 score on the corresponding development set (Dev-F1) used to identify the best model.

As can be seen in Table 1, the multi-task models (*MT*) consistently outperform the corresponding single-task models (*ST*) in terms of test F1 score, with the exception of the *Combined* set, when BERT is used a the base module. It is interesting to see that the models, however, have very little or no gain when only using identical tasks (*LMT*). On the other hand, we see the highest overall performance improvement of 14.76% for *MT* as compared to *ST*, on the *Combined* set, with *Albert* as shared module, and all the non-location tasks.

Table 2 shows the macro average of percentage improvement in F1-score from single-task models to the two variants of multi-task models for the three transformer-based modules. In general, we see significant performance improvement from the single-task to the multi-task model. From the table, Albert has the best overall performance improvement for the task of location identification. This could be because Albert has significantly less parameters compared to the other two variants, and given the small size of the datasets, that results in better generalization and shared learning. The overall lower performance gain from ST to LMT could be due to the fact that the two tasks are very similar and thus, they do not help model's generalization capability. On the other hand, when using more dissimilar tasks, the model benefits from variations in the data and is able to generalize better; this results in performance improvement across all variants.

Transformer	Average performance improvement (%)	
	ST $\rightarrow$ LMT	ST $\rightarrow$ MT
BERT	-0.78	3.43
RoBERTa	**4.75**	2.36
Albert	4.17	**6.35**

Table 2: Average (marcro) performance improvement over three datasets for the three transformer variants, when comparing single-task model (ST) to location-only multi-task (LMT) and all-multi-task (MT) model.

6 Conclusions and Future Work

In this paper, we studied the effect of using other closely related tasks in a multi-task setting for the task of fine-grained location identification in the domain of crisis-related tweets. Our results show that using multi-task learning can improve the performance significantly as compared to single-task learning. The approach could be specially useful when analysing an emergent crisis, where annotating large amount of data is not feasible and the entire analysis process is time-critical. Moreover, our results also show that when a small amount of data is available, the model may benefit more from using other related tasks rather than using additional data for the same type of task (i.e., two location identification tasks).

In addition to studying the effect of multi-task learning, we also studied the effect of different transformer variants on the performance of the model for the main task. With different variants available, it is necessary to find the best-fitting one for the task at hand. The results show that the choice of the shared module has significant effect on the performance.

As part of future work, we plan to explore other multi-tasking and domain adaptation strategies with different collections of datasets to allow for stress testing of the approach in terms of generalization. Given the promising results with multi-task learning, we plan to perform studies with respect to the nature of the crisis for the dataset. This will allow us to focus on specific datasets for specific crises and potentially improve the state-of-the-art on those datasets.

Acknowledgements

We thank the National Science Foundation and Amazon Web Services for support from grant IIS-1741345, which supported the research and the computation in this study.

References

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. *arXiv preprint arXiv:1805.00713*.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Bin Cheng, Guangcan Liu, Jingdong Wang, Zhongyang Huang, and Shuicheng Yan. 2011. Multi-task low-rank affinity pursuit for image segmentation. In *2011 International Conference on Computer Vision*, pages 2439–2446. IEEE.

Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2020. On identifying hashtags in disaster twitter data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 498–506.

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

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

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*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

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

Stuart E. Middleton, Lee Middleton, and Stefano Modafferi. 2014. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, 29(2):9–17.

Thai-Hoang Pham, Khai Mai, Nguyen Minh Trung, Nguyen Tuan Duc, Danushka Bolegala, Ryohei Sasano, and Satoshi Sekine. 2019. Multi-task learning with contextualized word representations for extented named entity recognition.

Jinfeng Rao, Ferhan Ture, and Jimmy Lin. 2018. Multi-task learning with neural networks for voice query understanding on an entertainment platform. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 636–645.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Daniel Stoller, Sebastian Ewert, and Simon Dixon. 2018. Jointly detecting and separating singing voice: A multi-task approach. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 329–339. Springer.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

T. Wang, Y. Tao, S. C. Chen, and M. L. Shyu. 2020. Multi-task multimodal learning for disaster situation assessment. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 209–212.

Yongxin Yang and Timothy M. Hospedales. 2017. Trace norm regularised deep multi-task learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Kiran Zahra, Muhammad Imran, and Frank O. Ostermann. 2020. Automatic identification of eyewitness messages on twitter during disasters. *Information Processing & Management*, 57(1):102107.

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *Computer Vision – ECCV 2014*, pages 94–108, Cham. Springer International Publishing.

A Hyperparameter configuration

Hyperparameter	**Values**
Learning rate	1e-5
Learning rate decay	Linear
Batch size	{**4**, 8, 16}
Classifier Dropout	0.1
Optimizer	AdamW
Adam epsilon	1e-8
Maximum Epochs	50

Table 3: Hyperparameter configuration used in the experiments. The best performing setting is highlighted in bold face.