Deep Multi-Task Learning to Extract Knowledge From Text

Alexander Hepburn

@ *a.hepburn.1@research.gla.ac.uk*

@alexjhepburn

University of Glasgow

University of Glasgow

Motivation

- Assisting crisis management/public health organisations by identifying information on natural disasters, emergencies, and pandemics on social media.
- Significant proportion (>10%) of the data we have collected contains information that would be considered useful to first responders. How do we accurately extract this actionable information?
- Twitter data is very noisy, short text is limited in how much information it gives us.
- How do we capture information about an event as it develops? How do we determine if the information learned is still relevant? How do we remove redundant information?
- Introduction to multi-task and incremental learning.

University of Glasgow

Multi-Task Learning

- Auxiliary, related tasks results in improved learned representations.
- Domain knowledge required to find tasks that are complementary, and reduce noise.
- Transformer architectures have resulted in SOTA results for a number of NLP tasks.

University of Glasgow

Training on co-related tasks can result in better performance

Example	Predictions on one task...	...can help disambiguate other tasks
X works for Y	RE: {*work, X, Y*}	*X* ⇝ Person (EMD) *Y* ⇝ Organization or Person (NER)
I love Melbourne. I've lived three years in this city.	CR: (*Melbourne, this city*) RE: {*live, I, this city*}	*Melbourne* ⇝ Location (EMD/NER)
Dell announced a $500M net loss. The company is near bankruptcy.	CR: (*Dell, The company*)	*Dell* ⇝ Organization (EMD/NER)

Fig. 1: Example of task correlation. "*A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks*", Sanh et. al (2018)

University of Glasgow

Why is this important?

- A lot of natural language processing relies on rich feature representations, lots of training data.
- Inductive transfer between tasks can alleviate challenges in low-resource settings.
- Tasks can be learned concurrently (multi-task), or a single task is learned using prior knowledge from training on previous, co-related tasks (transfer).

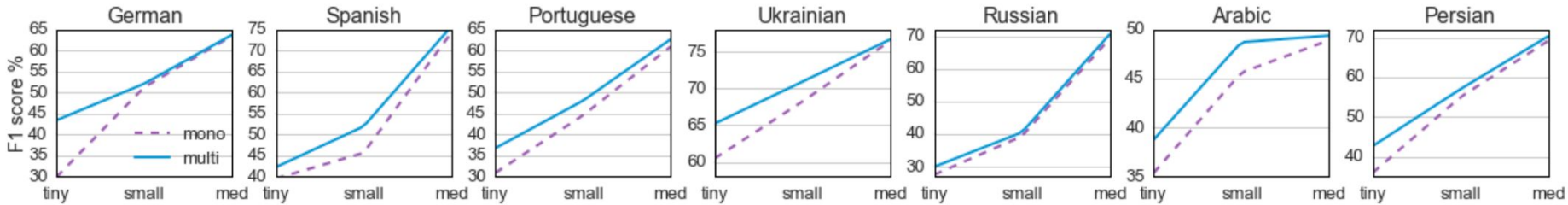


Fig. 2: F1 score comparison between monolingual and multilingual multi-task learning in low resource settings.“*Multilingual Hierarchical Attention Networks for Document Classification”*, Pappas et. al (2017)

How and why does it work?

- Hidden layers are shared between each task while maintaining task-specific output layers.
- Different tasks have different noise patterns, learning tasks simultaneously results in a more general representation.
- Relevant patterns between co-related tasks encourage the model to focus its attention on features that matter.
- Some features can be more complex to learn, introducing another task allows the model to **eavesdrop**, ie. learn task B through task A.

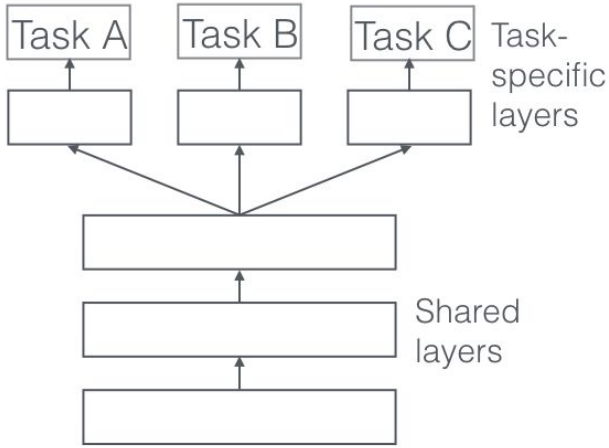


Fig. 3: Example of hard parameter sharing, *"An Overview of Multi-Task Learning in Deep Neural Networks", Ruder (2017).*

University of Glasgow

How does this relate to crisis management?

Information	Description	# Tweets	%
Terms/Phrases	Individual terms or phrases are information bearing, such as 'trapped' or 'lost power'.	164	97%
Location	The text explicitly mentions a location that is relevant to the event and the information contained is about that location	150	88%
Event Mention	The text explicitly mentions the event, making it easier to determine that this is relevant	34	20%

- Combining tasks such as part-of-speech tagging, named entity recognition, coreference resolution can result in our model learning a better understanding of terms/phrases and entities that define **actionable** information.

Regional Context Needed	To understand the tweet some additional information (not present in the tweet) is needed, such as an understanding of geographical landmarks in the affected area	42	25%
Tweet is Out of Date	The new assessor noted that based on the time-stamp of the tweet and when the information contained first became available, the information contained could be considered as out of date.	41	24%

- Deeper understanding of a word's representation can address some of the issues present in identifying important location information during crises.
- Alleviating the problem of few training examples can be useful in learning from unprecedented events (eg. COVID).

Fig. 4.1, 4.2: Features of information labelled as "critical priority", *"Incident Streams 2019: Actionable Insights and How to Find Them", McCreadie et al. (2020).*

University of Glasgow

Incremental Learning

- Can help us capture sequential elements in events.
- Moving away from traditional batch learning.
- Incremental introduction of tasks in an MTL framework have shown to outperform concurrent learning.

Impact of incremental learning

- Fairly new field, still a lot of research to be done. Also known as **online** or **continual learning**.
- Earlier work in effective text summarisation tasks have shown an almost double increase in performance when summaries were updated incrementally. (McCreadie et al., 2014)
- Baidu introduce ERNIE, incrementally introducing tasks in MTL frameworks which have shown to outperform SOTA model BERT over every task. (Sun et al., 2019)

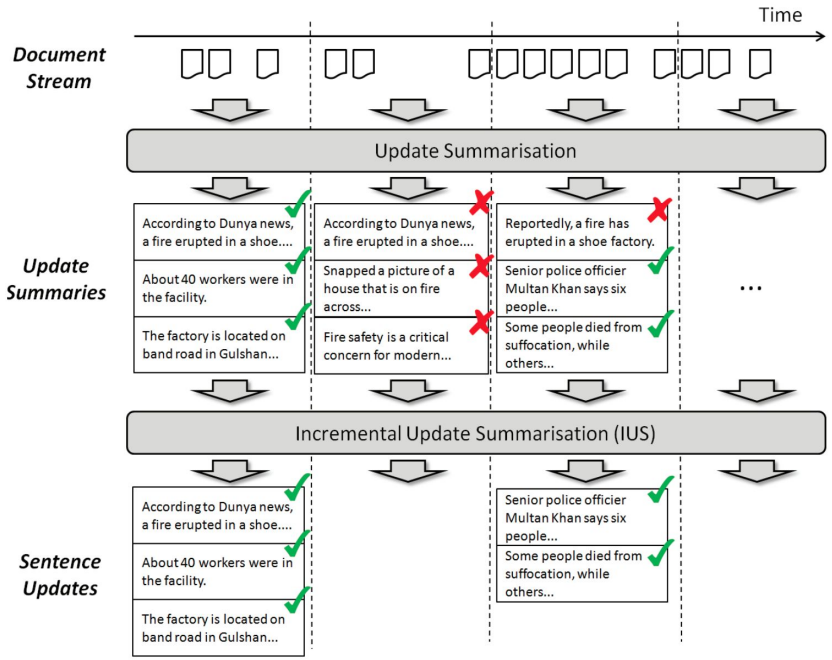


Fig. 5: Example of candidate summary selection in sequential updates.
"*Incremental Update Summarization: Adaptive Sentence Selection based on Prevalence and Novelty*", McCreadie et al. (2014)

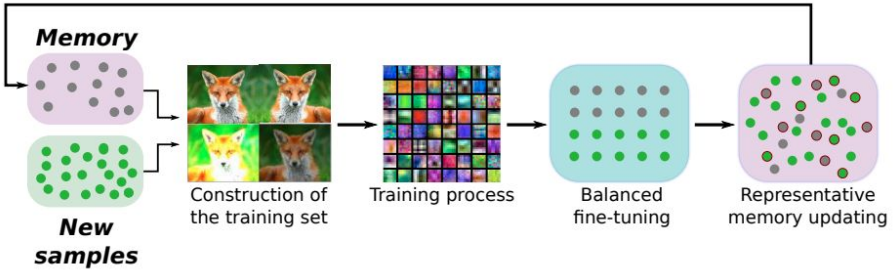
University of Glasgow

Why is this important?

- Our knowledge of an event changes greatly from start to finish, how do we reflect these updates in our learned representations?
- Traditionally models are trained in batches, what effect could incremental learning have in attaining the goal of (close to) real-time learning?
- MTL has proven to be greatly successful in text summarization tasks (Zhang et al., 2019), can we build on previous work in incremental learning to provide a **multi-task, incremental learning** framework?

University of Glasgow

Recent developments



- **OpenTapioca** (Delpeuch, 2019)* is a system which introduces the concept of real-time entity linking with knowledge base Wikidata.
- Packages such as **scikit-multiflow** (Montiel et al., 2018) have been developed to support machine learning for data streams.
- Castro et al., 2018 show a much slower degradation in the accuracy of their CNN model trained incrementally on ImageNet.

* "*OpenTapioca: Lightweight Entity Linking for Wikidata*", Delpeuch 2019 is a preprint and still under review.

University of Glasgow

Challenges

- Incremental learning is a very new field.
- Arbitrary number of labels to classify an event currently results in very poor classification results.
- Events can happen anywhere, some events very unlikely to happen in same location twice which results in little to no training data for regional information.
- **Catastrophic forgetting** means neural networks have a tendency to abruptly forget previously learned information with the introduction of new representations.

University of Glasgow

Where do we go from here?

- Regional, contextual information is crucial in identifying information for first responders:
  - ie. Exact location of a shooting, what hospital needs supplies of PPE, reports of food bank locations that need assistance.
  - How can we use external knowledge bases to enrich the limited information in short text?
- Can we use machine translation for non-English language tweets?
  - Is it worth examining multilingual models to account for events that are very prominent in particular regions of the world?
- Identify tasks or develop new, related tasks that can assist our MTL framework in understanding the context of a tweet.
- How do we eliminate redundant event knowledge in a model?
  - How can we further use these knowledge bases to update background information of, for instance, the current pandemic, as it develops?


University of Glasgow
Thank you for listening!
Any questions?
#UofGWorldChangers
@UofGlasgow