Initial Project Plan

Conversational Explanations for Deep Learning Systems

Cardiff University

Author: Oliver Schwertfeger

Student ID: 1403554

Supervisor: Professor Alun Preece

30th January 2018

Project Description

Over the past 10 years we have seen an exponential growth in data [1]. From generation to analytics, the use of 'Big Data' has transformed our world. The amount of data we generate has been growing substantially. In fact, by 2020 there is expected to be a 50-fold increase [1] in amount of data available since 2010. And that the amount of data stored on the internet will be around will be 40 zettabytes by 2020 [1].

This increasing amount of data from sources such as social media and IoT (Internet of Things) networks has made it infeasible for human analysts to analysis all the available data. This volume and variety of data, along with improving computational power, has made data-hungry algorithms like deep learning increasingly popular [2]. Deep learning models work, in short by artificially modelling the human brain. Given sufficient training data, these models can be incredibly effective at pattern matching and recognition compared to other classical linear classifiers. Recent applications of deep learning models include cancer prediction [3] and object recognition for autonomous cars [4]. However, little is known as to how a deep learning algorithm reaches its conclusion. We refer to this as the 'black box' of deep learning.

The motivation for this project comes from the lack of transparency of this 'black box'. The project aims to improve transparency of a deep learning-based classifier which, will classify social media data as 'interesting' in the domain of crime and security, providing its explanations for positive classification to the user through a conversational interface.

The project will be initially split into two parts. Part one is the creation of a deep learning-based classifier and part two is the addition of interpretability to the classifier.

Therefore, the primary goal of this project is to build an interpretable deep learning-based classifier.

Project Aims and Objectives

Listed below are the three main aims of the project.

Creation of a tweet classifier

The first aim of the project is to build a deep learning-based classifier that will be able to identify 'interesting tweets' from a dataset set given by the Cardiff University Crime and Research Institute. An 'interesting tweet' is a tweet from Twitter.com that has been flagged by human analyst which is considered to be worth analysing in the domain of crime and security.

The classifier will originally be created using unsupervised learning techniques, if this is unsuccessful then supervised learning techniques will be used.

The classifier should satisfy these criteria.

- The classifier should achieve a good level of accuracy once cross-validation is used.

The deliverables for aim this are as follows.

- **Demo** of the classifier to Friary House. Likely to be in week 6.
- **Deliverable** for the classifier is expected to be completed by 11th March.

Only after the system provides a high percentage of true positive results will this aim be considered complete.

Addition of conversational interface

Once it has been decided that the classifier is accurate, then the project will move onto its second aim, the addition of interpretability to the classifier. The idea behind adding interpretability is to aid a user in understanding as to how the classifier has reached its decision. So that the user can gain trust that the classifier is working correctly. The use of Mythos of Model Interpretability [5], will be used to influence the implementation of this aim.

- The interface should return in text format, the key words that have led to a tweet being classified as 'interesting'.
- The interface should handle errors well, if an unknown question is asked, useful feedback is given back.

The deliverables of this aim are as follows:

- **Demo** Working conversational prototype. Likely to be ready in week 9.
- **Deliverable** Conversational interface. Expected in week 10.

Addition of interactive conversational interface

The third aim of this project is to add an *interactive* element to the conversational interface. This would be achieved by using Amazon's Alexa for voice recognition. Allowing a user to verbally query the classifier, with the Alexa system relaying the reasoning behind the positive classification of a tweet. Subsequently, this would allow the classifier to be deployed in the Oscar environment at the Cardiff University Crime and Research Institute.

The deliverables of this aim are as follows:

- **Demo** Interactive conversational interface. Expected in week 12.
- **Deliverable** Interactive conversational interface. Expected to be completed in week 13.

This aim would be considered complete once a working prototype has been put implemented.

Milestone overview

Below is a list of monthly milestones I expect to achieve, with deliverables linked in. The smaller 'mid' milestones are to ensure progression throughout the project.

February:

Mid

- Be on track to achieve project milestone.
- Gain full understanding of project scope and what is required.

End

- Complete initial research and prototyping of creating a deep learning based classifier.
- Comfortable in the choice in deep learning libraries and tools.
- Have researched into method to add interpretability to the classifier. (Use of Myths of Model Interpretability [5]).

March:

Mid

- **Deliverable:** Achieve basic classifier prototype**.**
- Research of libraries such as TensorFlow with lime for interpretability.
- Have made good process on adding interpretability to the classifier.

End

- Decision if part two is feasible within the remaining time frame.
- Completed majority of the project research.
- Be on track with writing of final report.

April:

Mid

- Begin work on interactive conversational interface.
- Get feedback on how interface can be improved.
- Get outside testing of edge cases for conversational interface.
- **Deliverable:** Basic conversational interface for the classifier.

End

- **Demo:** Prototype interactive interface for the classifier.
- **Deliverable** Prototype interactive interface for the classifier.

May:

- Report finalisation.
- **Deliverable** Submission of final report and project code.

Ethics

After consulting the Cardiff University School of Computer Science Ethics guide [6], I believe that there is no ethical consideration required. This is because all identifiable data stored is publicly available (as Tweets via Twitter.com).

Work Plan

Week number	Dates	Plan
1	(29/01/18 – 04/02/18)	• Writing of initial project plan. • Background research into topic areas.
2	(5/02/18 – 11/02/18)	• Evaluate the existing DL models and libraries. • Research into methods to improve data quality. • Initial meeting with Alun Preece at Friary house. • Continue further reading into applications of text-based classification. • *Milestone* Gain full understanding of project scope and what is required. • **Deliverable** Submission of initial project plan. Due 5/02/18.
3	(*12/02/18 – 18/02/18)*	• Begin prototyping of classifier. • Investigation into generator and encoder components from Rationalising Neural Predictions [7] • Collection of data from Crime and Security Research Institute. • *Milestone* To be confident in choice of tools / library used. • Research into conversational interface development.
4 & 5	(19/02/18 – 4/03/18)	• Meeting with Alun Preece at Friary house. • Start work on classifier prototype using 'interesting tweets' dataset. • Continue to update draft final report.
6	(05/03/18 – 11/03/18)	• Meeting with Alun Preece at Friary house.

		• Feedback and refinement of classifier. • **Demo** Finalised classifier to team at Friary house by 9/03/18. • **Deliverable** Classifier due. Target 11th March.
7	(12/03/18 – 18/03/18*)*	• Begin work on conversational interface. • Refinement of the classifier (if needed) • Continue to update draft final report. • *Milestone* Gain further knowledge in adding interpretability.
8 & 9	(19/03/18 – 01/04/18)	• Meeting with Alun Preece at Friary House. • Easter recess begins 24th March. • **Demo** Basic conversational interface demo • *Milestone* Conclusion if all project aims are achievable in timeframe.
10 & 11	(02/04/18 – 15/04/18)	• Easter Recess. • **Deliverable** Basic conversational interface. Target due 10th April. • Begin development of interactive interface. (Alexa work starts). • Continue to update draft final report.
12	(16/04/18 – 22/04/18)	• Term restarts: 16th April. • Continued development and testing of interactive interface.
13	(23/04/18 – 29/04/18)	• Writing of final report. • Coding work likely to stop at this point. • **Demo** Final showcase of completed work to Friary House.

		• **Deliverable** Interactive conversational interface. Target due 29th April.
14 & 15	(30/04/18 - 11/05/18)	• Finalization of report. • Prepare for Viva. • **Deliverable** Submission of code and project report**.** Due 11th May.

Supervisor meetings

I currently have the following provisional meetings booked with Alun Preece.

- Monday 5th February 2018
- Monday 19th February 2018
- Monday 5th March 2018
- Monday 19th March 2018

Risk mitigation

The two-part nature of the project is well suited to avoiding risk. If part one is deemed to be more difficult than originally planned, then part two can be reduced or scrapped to give more time for completion of part one. Or if the classifier is considered unfeasible (for example, poor training data), then I would continue onto part two of the project. Creating a conversational interface for an existing system/tool instead.

Gant Chart

For illustrative purposes, I have an included a GANT chart to visually represent the work plan.

Week	1	2	3	4	5	6	7	8	9	10	11	12	13	14
W/C	29/01/2018	05/02/2018	12/02/2018	19/02/2018	26/02/2018	05/03/2018	12/03/2018	19/03/2018	26/03/2018	02/04/2018	09/04/2018	16/04/2018	23/04/2018	30/04/2018
Task														
Initial plan														
Research into classifier														
Classifier prototyping														
Creation of classifier														
Research of interpretability addition														
Implementation of interpretability addition														
Testing and feedback added interpretability														
Research of interactive interpretability														
Testing of interactive interpretability														
Final report writing														

References

[1] IDC Digital Universe Study: EMC, 30 September 2015. [Online]. Available: http://www.whizpr.be/upload/medialab/21/company/Media_Presentation_2012_DigiUniverseFINAL1.pdf. [Accessed 03 01 2018].

[2] O. E. Y. W. M. O. T. B. Hongming Chen, "Science Direct," 31 January 2018. [Online]. Available: https://ac.els-cdn.com/S1359644617303598/1-s2.0-S1359644617303598-main.pdf?_tid=7e3b7e30-092c-11e8-95a8-00000aab0f02&acdnat=1517694914_550107e75ec4090b3970d69b5bf55658. [Accessed 03 02 2018].

[3] K. . Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction.," , 2015. [Online]. Available: https://ncbi.nlm.nih.gov/pubmed/25750696. [Accessed 1 2 2018].

[4] Y. D. C. G. Ayşegül Uçar, "Object recognition and detection with deep learning for autonomous driving applications," 2 June 2017. [Online]. Available: http://journals.sagepub.com/doi/pdf/10.1177/0037549717709932. [Accessed 03 02 2018].

[5] Z. C. Lipton, "The Mythos of Model Interpretability," 6 March 2017. [Online]. Available: https://arxiv.org/pdf/1606.03490.pdf. [Accessed 03 02 2018].

[6] C. University, "Cardiff University Computer Science and Informatics Ethics," [Online]. Available: https://www.cs.cf.ac.uk/ethics/. [Accessed 03 02 2018].

[7] T. Lei, R. Barzilay and T. S. Jaakkola, "Rationalizing Neural Predictions," *arXiv: Computation and Language,* pp. 107-117, 2016.