DEPARTMENT OF ENGINEERING MATHEMATICS

Project Plan: Bayesian Deep Learning For Extractive Test Summarisation

James Stephenson

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Friday 29th July, 2022

Supervisor: Dr. Edwin Simpson

Contents

Abstract

Text summarisation is a valuable technique that allows for computational processing of documents, saving readers hours in manual processing. Users have different summary requirements; however, current extractive summarisation systems construct generic summaries that are not tailored to the user's needs. Asking users for feedback for summary improvement is one solution to combat this problem, yet this introduces an additional step of manual processing. Thus we look to minimise the amount of required user feedback.

This project will investigate the feasibility of applying newly-developed techniques from Bayesian deep learning [21] to get significant estimates of the model's confidence, so we are able to ask the user for more explanatory feedback. Legacy approaches use Bayesian optimisation [19] strategies the achieve minimal user feedback. However, this strategy is blocked since modern summarisation techniques involve deep neural networks. These models cannot effectively express uncertainty and are typically overconfident when encountered by new topics [22].

We will use existing exploratory frameworks [19] for evaluation.

Chapter 1

Two Pages of Coherent Text: An Introduction

It should include at least two pages of coherent text (i.e., of the form you intend to write for your thesis, not rough-notes) that could be used as the opening pages of your Introduction/Overview chapter (Chapter 1 of your thesis).

Chapter 2

Subject Background

Since the crux of this project is to assess the suitability of applying Bayesian deep learning (BDL) techniques to passage ranking (PR) problems, this chapter explores the relevant literature that discusses previous approaches to passage ranking solutions. Once this assessment has been done, we will then also examine literature that assesses BDL as opposed to classical deep learning techniques.

2.1 Interactive Learning

Interactive learning is a machine learning workflow involving directed experimentation with inputs and output [1]. Rapid change in response to user input facilitates interactive inspection on impact of user input. This workflow format is commonly used to solve NLP problems; related works include literature in passage ranking (PR) of generated text in the context of translations, question answering and text summarisation [12, 11, 14]. These works had a focus on interactionally-expensive uncertainty sampling to learn the rankings of *all* candidate passages [19]. Gao et al. [7] researched how to reduce the number of user interactions for uncertainty sampling techniques with some success using an active learner (AL). A positive step towards reasonable interactive learning.

Simpson et al [19] take an alternative approach to uncertainty sampling by proposing a Bayesian optimisation (BO) strategy instead [19]. With Gaussian process (GPs) displaying some success in error reduction for NLP tasks with noisy labels [3, 2], Simpson and Gurevych [20] proposed using Gaussian process preference learning (GPPL) with uncertainty sampling. This approach has been further built upon by Simpson et al. [19] to a BO framework. This approach showed a markable improvement in the accuracy of chosen answers in a community question answering (cQA) task with a small number of interactions required [19]. The methodology used Expected Improvement (IMP) as the acquisition function for AL which twisted the focus of optimisation to find the best candidate, as opposed to the ranks of all candidates [19]. The switch to exploitation of promising candidates showed to be massively influential on performance [19]. Simpson et al. [19] furthered the performance enhancement gained from using the BO framework by using prior predictions from a deep learner as an informative prior for GPPL [19] to address the cold-start problem for recommender systems [16].

2.2 Active Learning

Active

2.2.1 Active Learning Strategies

Active Strategies

2.3 Deep Learning

Deep learning methods form a subset of machine learning, based on neural networks with at least three hidden layers. These techniques have dramatically increased capabilities of model recognition in many domains including visual object recognition, question answering and text summarisation [10, 18, 23].

In classical training, one typically uses maximum a-posteriori (MAP) optimisation to choose the set of parameters, $\hat{w}$, for our model that maximises the posterior probability from our parameter distribution [21]. MAP does not require the computationally-costly calculations of the marginal distribution; however, since MAP is a point estimate, it cannot be fully considered a Bayesian approach [9].

2.3.1 Pre-trained Models

Pre-trained, deep learning, language models are useful in unsupervised learning problems due to the lack of major architectural modifications required and the high performance levels that are delivered [13]. One popular pre-trained language model is the Bidirectional Encoder Representations from Transformers (BERT) which takes an entire sequence of words at once to produce significantly improved results; with some minor fine-tuning, it can be applied to many NLP tasks [13].

Recent publications have found BERT-based models [4] to be extremely effective when tasked with passage ranking situations across the question answering and text summarisation disciplines [22, 15]. Xu et al. [22] explored a query-passage set up when applying BERT to cQA such that the BERT final hidden state fed into an MLP module to produce relevance scores in a supervised way. Since this technique outperformed the baseline, it may be a useful structure to consider adapting to the text summarisation domain.

The limitation of utilising an interactive learning framework such as one outlined by Simpson et al. [19] as that it does not utilise the vast performance capabilities of newer, pre-trained techniques such as BERT. Although the framework presented does limit the number of interactions required from a user – allowing the user to tailor the summary – Ein-Dor et al. [5] look to take this idea further with the incorporation of a BERT component in an AL framework.

2.4 Bayesian Optimisation

Bayesian Optimisation

2.5 Deep Active Learning

Ein-Dor et al. [5] developed a framework that used an AL approach with BERT-based classification. Zhang and Zhang also explored an alike ensemble of AL strategies [24]; however, the task is less relatable to PR since framework proposed by Ein-Dor et al. had experimentation on data with high class imbalance, scarce labelling and a small annotation budget [5], attributes of an interactive PR context. This structure consisted of pool-based AL [17] in batch mode in conjunction with BERT as the classification scheme. Different AL strategies were examined – Monte-Carlo Dropout (MCD) [6], a Bayesian approach, and Discriminative Active Learning (DAL) [8] – with Al proving an excellent boost to helping BERT emerge from its poor initial model [5]. Although DAL would not be appropriate for the PR context due to its focus on querying batches, using MCD as a strategy is a technique we could consider.

2.6 Bayesian Deep Learning

Deep bayes

2.6.1 Bayesian Deep Learning Strategies

Deep Bayes Strategies

Bibliography

[1] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, December 2014.

[2] Daniel Beck, Trevor Cohn, and Lucia Specia. Joint emotion analysis via multi-task Gaussian processes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1798–1803, Doha, Qatar, October 2014. Association for Computational Linguistics.

[3] Trevor Cohn and Lucia Specia. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[5] Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online, November 2020. Association for Computational Linguistics.

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2015.

[7] Yang Gao, Christian M. Meyer, and Iryna Gurevych. APRIL: Interactively learning to summarise by combining active preference learning and reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4120–4130, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[8] Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning, 2019.

[9] Alfred O. Hero. Statistical methods for signal processing. In *STATISTICAL METHODS FOR SIGNAL PROCESSING*, 2005.

[10] Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[11] Xiao Lin and Devi Parikh. Active learning for visual question answering: An empirical study, 2017.

[12] Álvaro Peris and Francisco Casacuberta. Active learning for interactive neural machine translation of data streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 151–160, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[13] M. Ph. D., Aklima Lima, Kamruddin Nur, Sujoy Das, Mahmud Hasan, and Muhammad Kabir. A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, PP:1–1, 11 2021.

[14] Avinesh P.V.S and Christian M. Meyer. Joint optimization of user-desired content in multi-document summaries by learning from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1353–1363, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[15] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. Understanding the behaviors of bert in ranking, 2019.

[16] Jesus Bobadilla Sancho, Fernando Ortega Requena, Antonio Hernando Esteban, and Jesús Bernal Bermúdez. A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems*, 26:225–238, February 2012.

[17] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[18] Yashvardhan Sharma and Sahil Gupta. Deep learning approaches for question answering system. *Procedia computer science*, 132:785–794, 2018.

[19] Edwin Simpson, Yang Gao, and Iryna Gurevych. Interactive text ranking with bayesian optimisation: A case study on community qa and summarisation, November 2019.

[20] Edwin Simpson and Iryna Gurevych. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371, 2018.

[21] Andrew Gordon Wilson. The case for bayesian deep learning, January 2020.

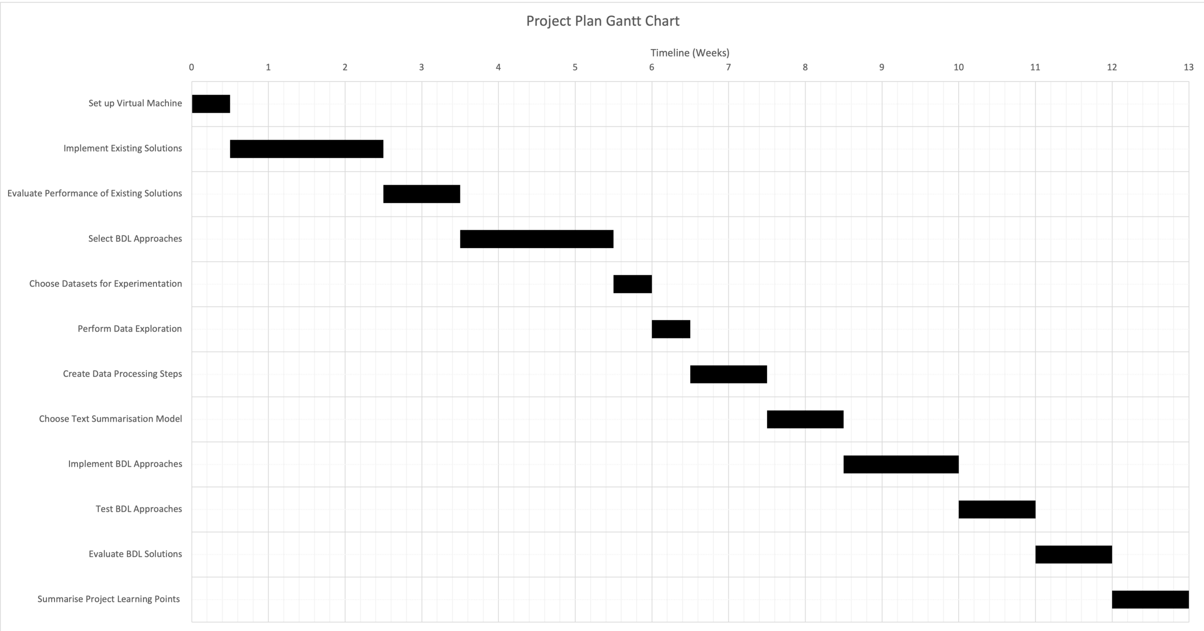
[22] Peng Xu, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Passage ranking with weak supervision. *May*, 2019.

[23] Mahmood Yousefi-Azar and Len Hamey. Text summarization using unsupervised deep learning. *Expert Systems with Applications*, 68:93–105, 2017.

[24] Leihan Zhang and Le Zhang. An ensemble deep active learning method for intent classification. In *An Ensemble Deep Active Learning Method for Intent Classification*, pages 107–111, 12 2019.

Appendix A

Time-plan



Appendix B

One-Page Risk Assessment

It should include as an Appendix a one-page risk assessment for your project, talking about the major risks you can foresee that might plausibly occur and interfere with your plans. For each risk, state clearly what it is, what its likelihood is, what its effects/impact would be on the project, and what your intended mitigation or risk-reduction involves.