Literature Review

*Intro*

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.

*Previous works*

* Interactive Learning

Interactive learning is a machine learning workflow involving directed experimentation with inputs and output \cite{Amershi14}. 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 \cite{Peris18}, \cite{Lin17}, \cite{PVS17}. These works had a focus on interactionally-expensive uncertainty sampling to learn the rankings of \emph{all} candidate passages \cite{Simpson19}. Gao et al. \cite{Gao18} researched how to reduce the number of user interactions for uncertainty sampling techniques with some success using an active learner. A positive step towards reasonable interactive learning.

Simpson et al \cite{Simpson19} take an alternative approach to uncertainty sampling by proposing a Bayesian optimisation (BO) strategy instead \cite{Simpson19}. With Gaussian process (GPs) displaying some success in error reduction for NLP tasks with noisy labels \cite{Cohn13, Beck14}, Simpson and Gurevych \cite{Simpson18} proposed using Gaussian process preference learning (GPPL) with uncertainty sampling. This approach has been further built upon by Simpson et al. \cite{Simpson19} 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 \cite{Simpson19}. 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 \cite{Simpson19}. The switch to exploitation of promising candidates showed to be massively influential on performance \cite{Simpson19}. Simpson et al. \cite{Simpson19} furthered the performance enhancement gained from using the BO framework by using prior predictions from a deep learner as an informative prior for GPPL \cite{Simpson19} to address the cold-start problem for recommender systems \cite{Bobadilla12}.

* Active Learning

*What is AL?*

Alongside unsupervised and supervised learning, active learning (AL) is a machine learning framework whereby queries are asked of an oracle – such as a human annotator – in the form of labelling unlabelled observations \cite{Settles09}. The active interactions with oracles allow better performance with few labelled data points. AL is beneficial in the cases where labelled data is scarce due to high costs; for speech recognition problems \cite{Zhu05} details a scale factor of ten times between the length of a speech extract and the time taken to annotate such as extract.

*AL Scenarios*

Settles \cite{Settles09} describes three scenarios that are considered in literature to categories AL problems.

\paragraph{Membership query synthesis.}

\paragraph{Stream-based selective sampling.}

\paragraph{Pool-based active learning.}

*Acquisition Functions*

*Current Approaches*

* DEEP LEARNING

*What is 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 \cite{Lecun15}, \cite{Sharma18}, \cite{Azar17}.

[NN Deep Learning Image]

*MAP – what is done classically.*

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 \cite{Wilson20}. MAP does not require the computationally-costly calculation of the marginal distribution \cite{Hero14}; however, since MAP is a point estimate, it cannot be fully considered a Bayesian approach \cite{Hero15}.

[Hero15 plot]

* + Pre-trained Models - BERT

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 \cite{Mridha19}.

*What is BERT?*

One popular pre-trained language model is the Bidirectional Encoder Representations from Transformers (BERT) which takes an entire sequence of words, bidirectionally, to produce significantly improved results. The input is augmented by three embeddings – position, segment and token embeddings – and padded by a [CLS] token at the beginning of the first sentence to ensure BERT has lots of useful information \cite{Navin21}. BERT is trained on two tasks in parallel: Masked Language Modelling, prediction of hidden words in sentences, and Next Sentence Prediction \cite{Navin21}. However BERT can be applied to many NLP tasks \cite{Mridha19} such as question answering and text classification tasks with some minor fine-tuning; we add an additional small layer on the top of the transformer output for the [CLS] token \cite{Navin21} to adapt the core model to different tasks.

[classification example]

*How does it do?*

Recent publications have found BERT-based models \cite{Devlin18} to be extremely effective when tasked with passage ranking situations across the question answering and text summarisation disciplines \cite{Xu19}, \cite{Qiao19}. Xu et al. \cite{Xu19} 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. \cite{Simpson19} 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. \cite{EinDor20} look to take this idea further with the incorporation of a BERT component in an AL framework.

* Deep Active Learning

Ein-Dor et al. \cite{EinDor20} developed a framework that used an AL approach with BERT-based classification. Zhang and Zhang also explored an alike ensemble of AL strategies \cite{Zhang19}; 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 \cite{EinDor20}, attributes of an interactive PR context. This structure consisted of pool-based AL \cite{Settles09} in batch mode in conjunction with BERT as the classification scheme. Different AL strategies were examined – Monte-Carlo Dropout (MCD) \cite{Gal15}, a Bayesian approach, and Discriminative Active Learning (DAL) \cite{Gissin19} – with Al proving an excellent boost to helping BERT emerge from its poor initial model \cite{EinDor19}. 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.

* *Bayesian Deep Learning*

*What is BDL?*

Wilson and Izmailov \cite{Izmailov20} provide a good overview of

*Different Strategies*