PLMs- Pre-trained language models

1. **PASSAGE RE-RANKING WITH BERT:**

A simple re-implementation of BERT for query-based passage re-ranking. A simple Question-Answer pipeline consists of three main stages-

* 1. a large number (e.g., 1k) of possibly relevant documents to a given question are retrieved from a corpus by a standard mechanism, such as BM25.
  2. passage re-ranking - by a more computationally-intensive method
  3. top 10 or 50 of these documents will be the candidate answers

This paper discusses the second stage of the pipeline

**Method:**

Using **BERTLARGE** model as a **binary classification model**, estimate a score si of how relevant a candidate passage di is to a query q. We start training from a pre-trained BERT model and fine-tune it to our re-ranking task using the cross-entropy loss.

Input-

* 1. **Query**- sentence A (truncate up to at most 64 tokens)
  2. **Passage text**- sentence B (truncate passage such that concatenation of [query + passage text + separator tokens] have max length of most 512 tokens)

The probability for each passage is calculated independently. Final list of passages are obtained by ranking them with respect to these probabilities.

**RESULTS:**

**Training size vs performance**: The pretrained models require few training examples from the end task to achieve a good performance

1. **PARADE: Passage Representation Aggregation for Document Reranking**

There are many types of relevance signals that can only be observed in a full document.

It is not necessarily possible to account for all such excerpts by considering only the top passages. Similarly, the ordering of passages itself may affect a document’s relevance. Also, more relevant documents also contain a higher number of relevant passages

**Aggregating passage-level relevance scores** to predict the document’s relevance score outperforms the common practice of using the maximum passage score. On the other hand, the amount of non-relevant information in a document can also be a signal. Adding non-relevant information to a document should decrease its score.

It is found that **aggregation over passage representations** using architectures like CNNs and transformers outperforms **passage score aggregation.**

An analysis of following-

* 1. how to **reduce the computational cost** of transformer-based representation aggregation by decreasing the model size **[**using knowledge distillation**]**
  2. role of **number of passages considered** influencing the effectiveness of transformer-based representation aggregation
  3. **dataset characteristics** [influencing which aggregation strategies are best on certain benchmarks].

1. **Distilling Dense Representations for Ranking using Tightly-Coupled Teachers**
2. **The Role of Complex NLP in Transformers for Text Ranking?**

Syntax do not play a critical role in the effectiveness of re-ranking with BERT. The Cross-Encoder ranker further suggests that through fine-tuning the ranking task the sensitivity to the input order is weakened. The main attributions for the superior performance of BERT are-

* 1. query-passage cross-attention or deep matching
  2. richer embeddings (capturing word meanings based on aggregated context where word order has only a negligible influence)

Findings of this paper provide a good starting point for the design of pre-training tasks tailored specifically to ranking, possibly reducing complexity and model size.

1. **BERT Rankers are Brittle: a Study using Adversarial Document Perturbations**

BERT Rankers are applied in many texts ranking tasks with a huge performance gain and it comes at the cost of vulnerability against adversarial attack, as well as the lack of interpretability leading to wrong predictions.

Using gradient-based optimization methods, to search candidate tokens such that we can add or replace a small number of tokens to-

* 1. a highly relevant to cause a large rank demotion
  2. a non-relevant document to cause a large rank promotion.

The scope of an adversarial ranking attack can be-

* 1. local (on a given query): perturb retrieved document given a query.
  2. global (on an entire query workload): adversarial tokens can be generated for an entire workload of queries.

**Adversarial Attacks in NLP:** There are two attacking approaches-

* 1. black-box attack: the attacker uses predefined heuristic rules to generate natural substitutes of words or sentences to fool the victim model.
  2. white-box attack: full access to model parameters.

**Universal Triggers:** Triggers are phrases or tokens injecting which may mislead the ranking model to target prediction. There exist some tokens that always result in particular prediction in high dimensional neural networks. It is also possible to plant such triggers by adding a small set of crafted instances having such triggers in the training data.

1. **TILDE: Term Independent Likelihood moDEl for Passage Re-ranking:**

BERT-based, Term Independent Likelihood moDEl (TILDE), which ranks documents by both query and document likelihood. The transformer model’s inference step is computationally expensive and query likelihood cannot be pre-computed offline. TILDE achieves competitive effectiveness along with low query latency.

When TILDE is set to rely only on query likelihood (TILDE-QL), at query time the deep language model’s tokenization step (very fast operation) is the only additional step required to conventional inverted index retrieval.

When TILDE is set to rely on both query and document likelihoods (TILDE-QDL), then at query time both the deep language model’s tokenization and one typical deep language model’s inference step are both required: while reducing effectiveness, the inference step does provide additional relevance signals to the re-ranker, thus further increasing effectiveness.

Topics-

1. BERT rankers might automatically associate higher relevance to the terms present at the beginning of the document.

2. Genre classification of Documents using BERT