SPECTER Review : Document-level Representation Learning using Citation-informed Transformers

Syed Ali Shah
Data Science Engineering
Politecnico di Torino
S292423@studenti.polito.it

Bilal Shabbir
Data Science Engineering
Politecnico di Torino
S305979@studenti.polito.it

***Abstract*—This project presents an overview of the paper titled "SPECTER: Document-level Representation Learning using Citation-informed Transformers"[1]. The paper introduces SPECTER, a novel approach for document-level representation learning that leverages citation information to enhance the performance of transformer-based models. The report provides a summary of the key contributions, methodology, experimental results, and implications of the proposed SPECTER model. As an extension, We tried to improve SPECTER classification performance on the MeSH and MAG tasks by mounting a MLP classification head on top of SPECTER. The Code is available in the following github repository.**

I. Introduction

With the recent scientific advancements,Natural Language Processing (NLP) tools that aid users in searching, discovering, and understanding are becoming more and more important. Pretrained neural language models (LMs) have significantly improved NLP tools in recent years [3][4][5]. Although these models are frequently employed to represent specific words or phrases, applications to whole-document embeddings are yet mostly unexplored. The latest pretrained LMs have not yet been incorporated into approaches that do use inter-document signals to create whole-document embeddings [6][7]. The original specter paper investigates how to learn embeddings for scientific publications using the power of pretrained language models.

The title and abstract of a paper provide extensive semantic information about the paper, but as the paper reveals, passing these textual fields to even a cutting-edge model specifically designed for scientific text like the recent SciBERT [2] does not produce accurate paper representations. The model was pretrained using language modeling objectives, but these objectives did not produce output representations that were useful for tasks at the document level, including subject categorization or recommendation.

The specter model offers a novel technique for acquiring all-purpose vector representations of scientific texts. Without the need for any task-specific fine-tuning of the pretrained language model, The model incorporates inter-document context into the Transformer [9] language models, such as SciBERT [2], to learn document representations that are effective across a wide-variety of downstream tasks. Citations are especially used as a naturally occurring triplet-loss pretraining target that indicates which documents are more connected between them. Our model, in contrast to many earlier efforts, does not require any citation information at the moment of inference.

The authors of SPECTER use an SVM classifier to classify the embeddings and output the expected label. The MeSH and MAG classification process was changed as variation on top of SPECTER, and we used a variety of classification heads to correctly identify the embeddings.

II. The SciBERT model

The SPECTER model uses SciBERT [2] encoder as backbone. The SciBERT encoder is infact a BERT model[8] trained on scientific literature. It was trained using a corpus comprised of more than 1,000,000 full-text articles that was downloaded from semanticscholar.org. The authors have created an ad hoc vocabulary they term scivocab to better fit the training corpus. In tasks involving text classification on the SciCite, PubMed, and ACL-ARC datasets, SciBERT achieved state of art performances. SciBERT's performances on document-level tasks, such as the classification of scientific papers, are not as promising as those on the sentence-level tasks that we just mentioned, despite being very promising nonetheless due to its training procedure (which is designed to resemble BERT's).

III. SPECTER

As mentioned earlier SPECTER is based on SciBERT. By introducing a signal of inter-document relatedness into the model that is collected from the data in the citation network of each publication P, SPECTER overcomes the constraints of SciBERT. By creating triplets from the metadata of a query (PQ), a positive (P+), and a negative (P) paper, this signal is collected. Negative papers are those that are not cited by the query paper but may or may not be cited by the positive paper. This information allows us to distinguish between hard and easy negatives, respectively. Positive papers are those that are cited by the query paper.Then, a SciBERT embedder that has been trained to minimize the following loss function is fed these training triplets.

$$L = \max\{d(PQ, P^+) - d(PQ, P^-) + m, 0\} \qquad (1)$$

$d(P^A, P^B) = \|v_A - v_B\|_2$ is a measure of the difference between two distinct embeddings,where v is the pooled output of the SciBERT encoder. This loss clearly motivates the model to embed the query and the positive paper similarly while

differentiating the embedding of the query and the negative paper.

IV. Proposed Strategy

We propose to introduce neural network layers on top of the SPECTER embeddings for the classification task. For this purpose we experimented with three different setting. Using a single layers. Then we used classifier netwrork consisting of a hidden layer. We allso experimented with classifier made up of a network of 5 hidden layers. Aim of introducing adhoc network is to improve on the downstream classification task. Each hidden layer is formed by 64 units and all the classification heads do use the ReLU activation function. The following figure reaveals the architecture in a nutshell.

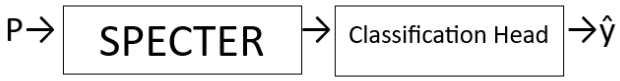

Fig. 1. Proposed Strategy

V. Experiments

Due to computational constraints, we used only a fraction of the MeSH and MAG databases . We specifically screened out papers that did not include an abstract or both a title, in addition to papers that were not written in English. The dataset was decreased by 24% through filtering. The remaining publications' titles and abstracts were then reduced, concatenated, and tokenized to form the input feature for the SPECTER embedder (with a [SEP] token in between). The dataset is now ready to be utilized for fine-tuning once a column containing the labels is added to it from auxiliary files. We specifically adjusted SPECTER for MeSH and MAG classification. The MeSH subset we used consists of 23154 labeled examples whereas the MAG subset consists of 14083.

Given the extent of the datasets under consideration, papers have used 90% to 10% holdout sets before dividing the datasets into train and test sets. After the papers were fed into the conventional SPECTER model, an MLP classifier would receive the embeddings corresponding to the typical [CLS] token. Only 5 epochs of the fine-tuning phase were run on each datasets using the two different models. on a batch size of 5. With a constant learning rate of 5e-5, we employed the AdamW optimizer and the Cross-Entropy Loss as our classification loss. We performed testing on Google Colab Tesla T4 GPUs. The following table shows the results obtained from classification using the above mentioned architecture. The results look very promissing and show improvements as compared to that of original SpPECTER model. However as the complexity of the network grows. Network requires more training to learn from the embeddings. Hence the results from the 5 hidden layer model seem downgraded.

Model	MeSH -F1	MAG-F1
SPECTER	0.87	0.79
SPECTER + Linear Layer	0.93	0.94
SPECTER + 1 hidden layer	0.96	0.93
SPECTER + 5 hidden layer	0.04	0.01

Fig. 2. Test results

VI. Conclusion

In conclusion, our comprehensive review of the Specter model highlights its impressive capabilities and potential impact in the realm of text summarization. Specter, built on a Transformer-based architecture and trained on extensive textual data, showcases a profound understanding of document context, enabling it to distill crucial information into concise summaries while preserving the essence of the original content. By leveraging self-supervised and transfer learning, Specter acquires a broad range of linguistic knowledge, making it adaptable to various domains and text genres. Its competitive performance against benchmark metrics demonstrates its effectiveness compared to other state-of-the-art summarization models. Specter's advancements have significant implications, offering the promise of efficient information retrieval, knowledge extraction, and text comprehension amidst the overwhelming influx of textual data. However, further research is needed to address challenges such as complex sentence structures and topic diversity. Specter's remarkable capabilities position it as a pivotal contributor to the field, driving the future of accurate and efficient text summarization in our information-driven society.

References

[1] Arman Cohan.Iz Beltagy Sergey Feldman Doug Downey Daniel S.Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers

[2] Iz Beltagy, Arman Cohan, and Kyle Lo. "SciBERT: Pretrained Contextualized Embeddings for Scientific Text".

[3] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. arXiv.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.

[5] Wenlin Wang, Chenyang Tao, Zhe Gan, Guoyin Wang, Liqun Chen, Xinyuan Zhang, Ruiyi Zhang, Qian Yang, Ricardo Henao, and Lawrence Carin. 2019. Improving textual network learning with variational homophilic embeddings. In Advances in Neural Information Processing Systems, pages 2074–2085.

[6] Cunchao Tu, Han Liu, Zhiyuan Liu, and Maosong Sun. 2017. Cane: Context-aware network embedding for relation modeling. In ACL.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar S. Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.

[8] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: CoRR abs/1810.04805 (2018).ù

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In NIPS.