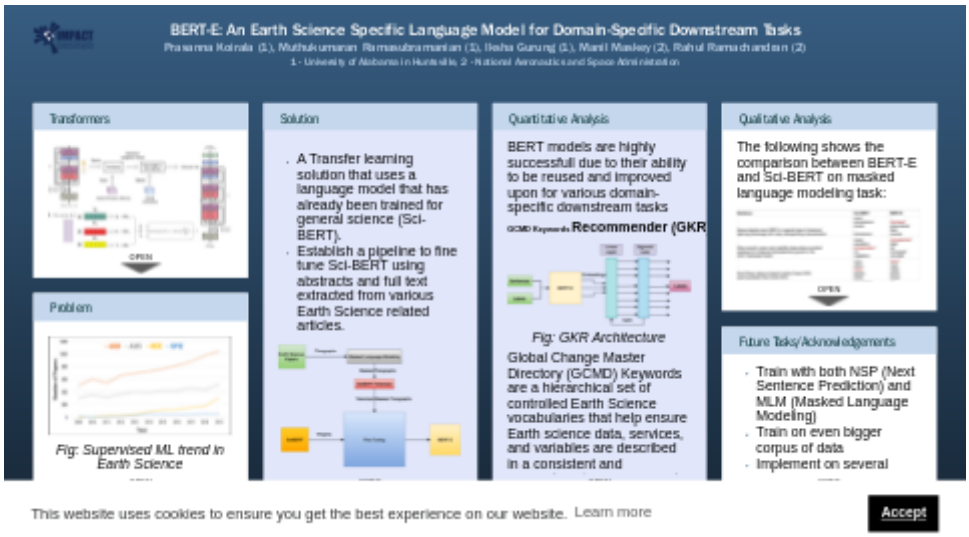
BERT-E: An Earth Science Specific Language Model for Domain-Specific Downstream Tasks



Prasanna Koirala (1), Muthukumaran Ramasubramanian (1), Iksha Gurung (1), Manil Maskey (2), Rahul Ramachandran (2)

1 - University of Alabama in Huntsville, 2 - National Aeronautics and Space Administration

PRESENTED AT:

AGU FALL MEETING
New Orleans, LA & Online Everywhere
13–17 December 2021
Poster Gallery brought to you by WILEY

TRANSFORMERS

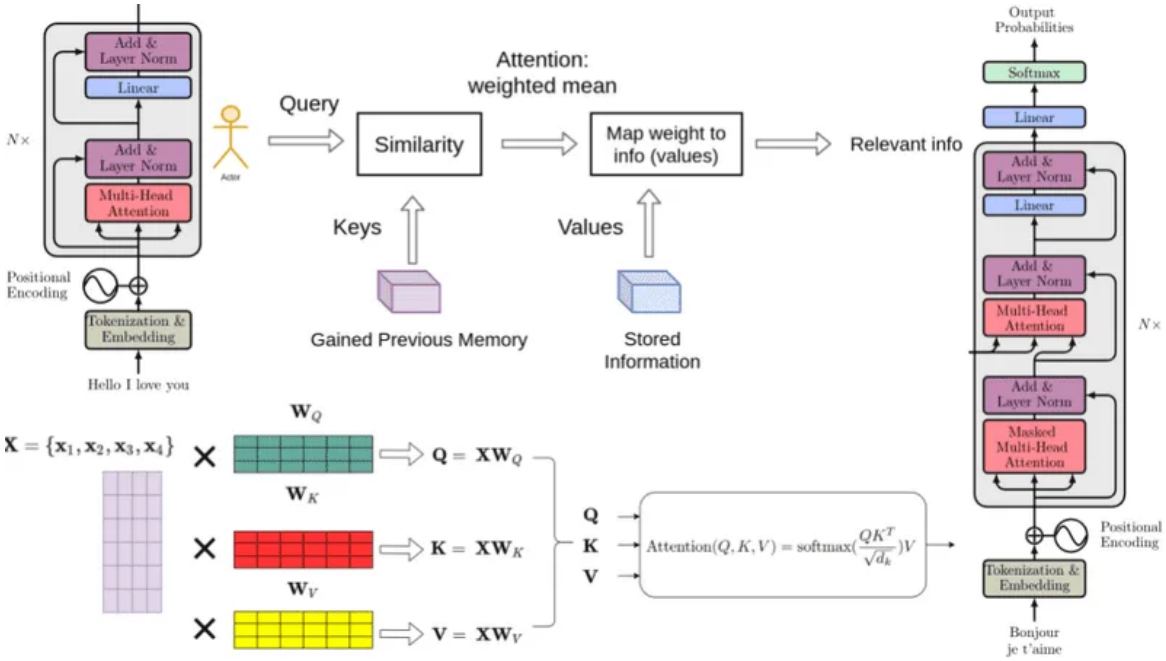


Fig: Architecture of Transformers
Source: https://theaisummer.com/transformer/

Transformers are machine learning language models trained on huge volumes of dataset using millions of parameters. Recent advances in Machine Learning has enabled transformers to perform with near human capabilities in a myriad of language tasks such as next sentence prediction, question answering, entity extraction and finding analogies.

Transformers were first introduced in a ground breaking paper titled "Attention is all you need"[1] in 2017. The paper talks about assiging attention to each word in a sentence in order to translate it to a different language.

Bidirectional Encoder Representations from Transformers (BERT) [2] was introduced in 2019 by researchers Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT is especially effective because it is understands the context of a word both left and right side of the sentence. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

PROBLEM

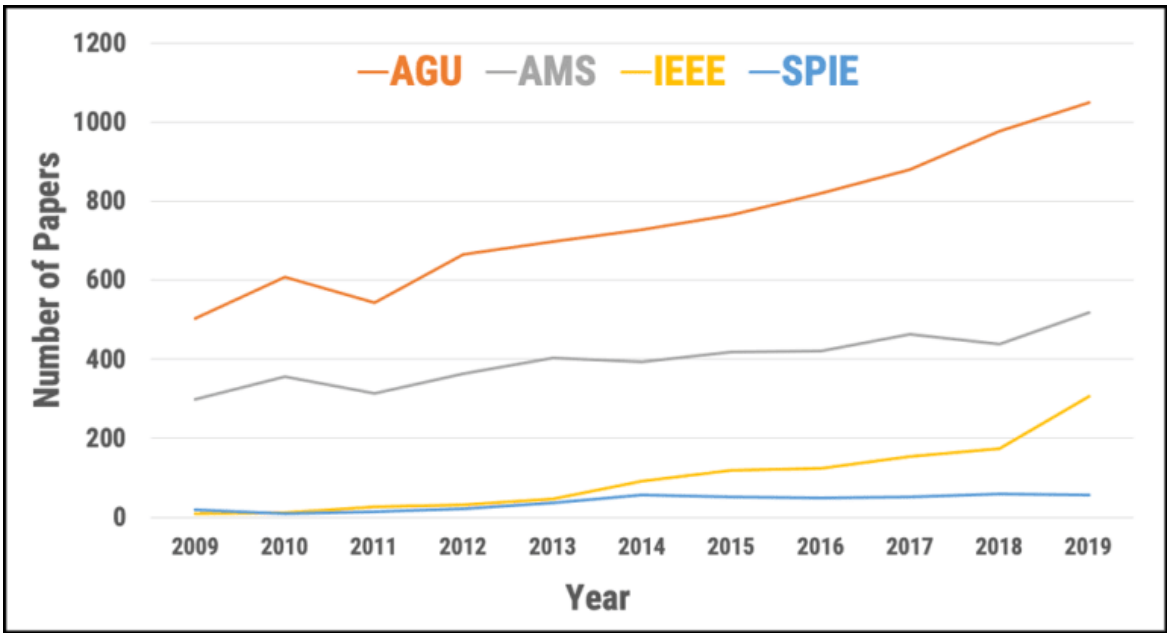


Fig: Supervised ML trend in Earth Science

- Even though there is an increasing trend in use of Machine Learning in the Earth science community, the practice is still fairly new to the community [3]
- Language models are specifically useful in the domain of Earth Science for tasks like contextual dataset keyword tagging using text metadata, question answering, relevant document retreival and domain entity extraction. However, language models need extremely large number of dataset for training. Procuring this dataset is a challenging task.
- Transformers are trained on varied natural language text over the internet. Hence, they work well for generalized case but not equally as well for domain specific tasks. Hence, there is a lot of room for improvement when using these models for earth science specific tasks

SOLUTION

- A Transfer learning solution that uses a language model that has already been trained for general science (Sci-BERT).
- Establish a pipeline to fine tune Sci-BERT using abstracts and full text extracted from various Earth Science related articles.

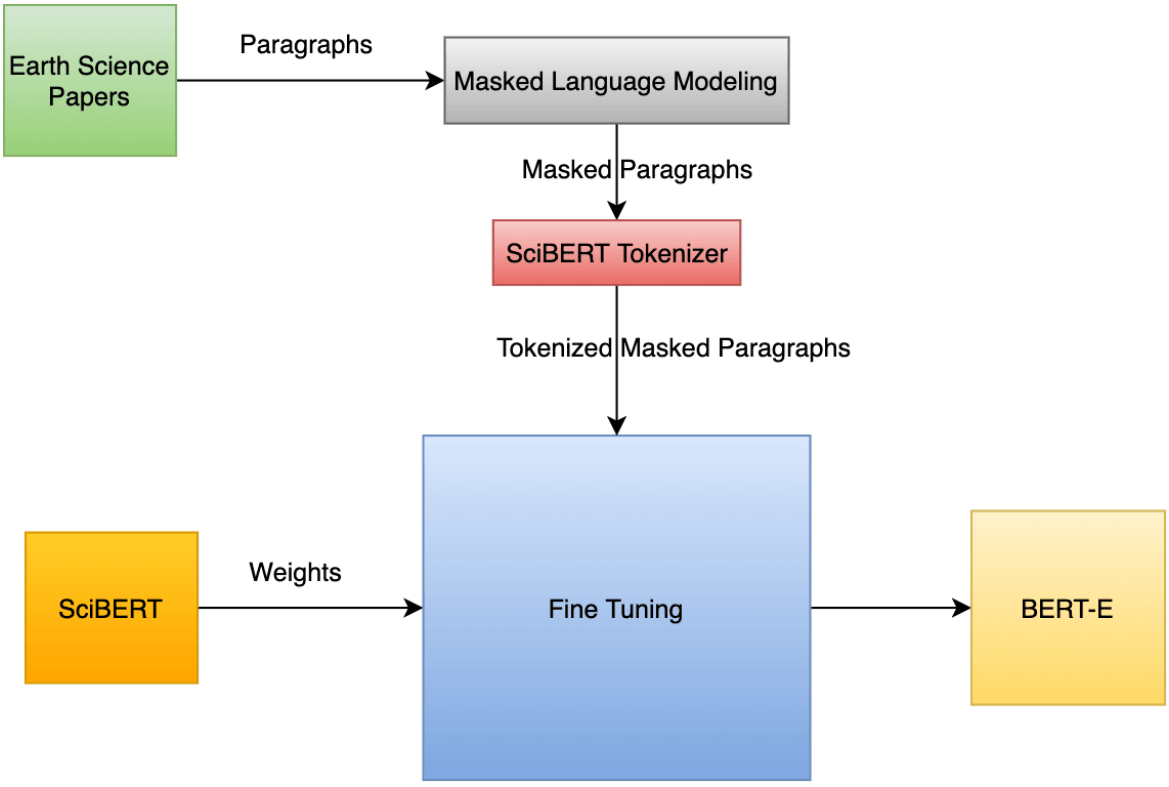


Fig: BERT-E architecture

Training Process:

- 270k+ Earth science articles with a total of almost 6 million paragraphs.
- The train eval split choosen was 95% and 5% respectively.
- The dataset is prepared into separate paragraphs which the Sci-BERT tokenizer tokenizes to form encodings.
- Masked Language Modeling (MLM) is used to mask random words in the paragraph. The model trains by predicting the word and validating againts the actual word.
- The training changes the model weights of the Sci-BERT model (fine tunes) such that the model hence generated is customized for the Earth science domain.

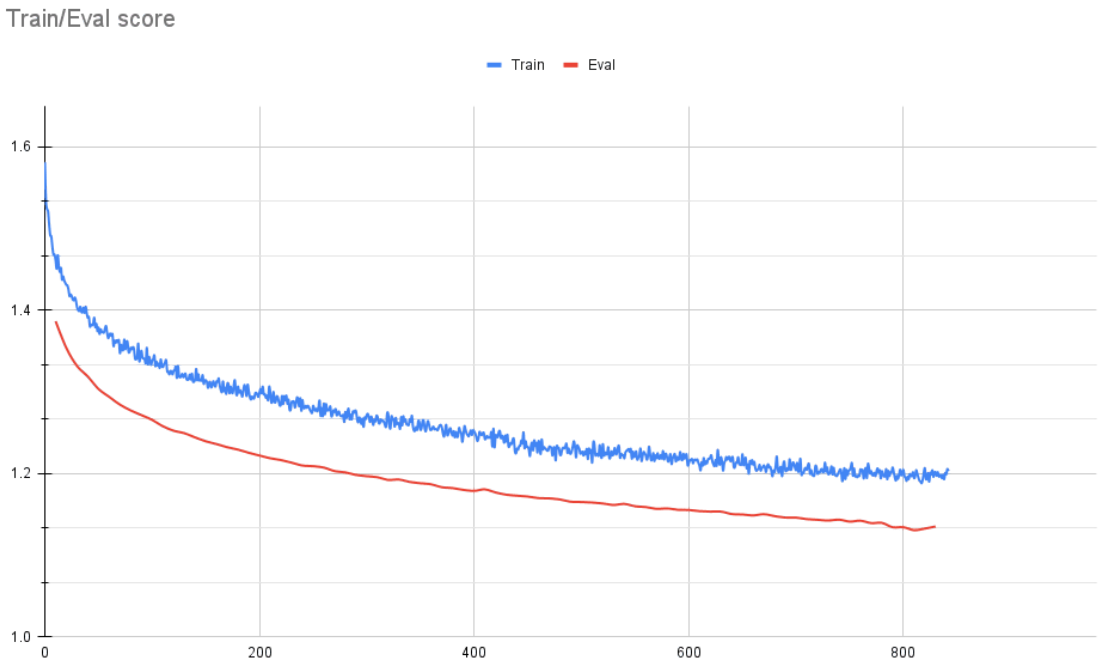


Fig: Train eval graph for the training

The eval loss is seen to be constantly decreasing and less than the training loss. This can be caused due to various factors, such as, eval data being "easy" as compaired to the training data and/or not being a good representation of the dataset. However, this was a calculated risk given resource constraints.

QUANTITATIVE ANALYSIS

BERT models are highly successfull due to their ability to be reused and improved upon for various domain-specific downstream tasks that regular language models dont do particularly well at.

GCMD Keywords Recommender (GKR)

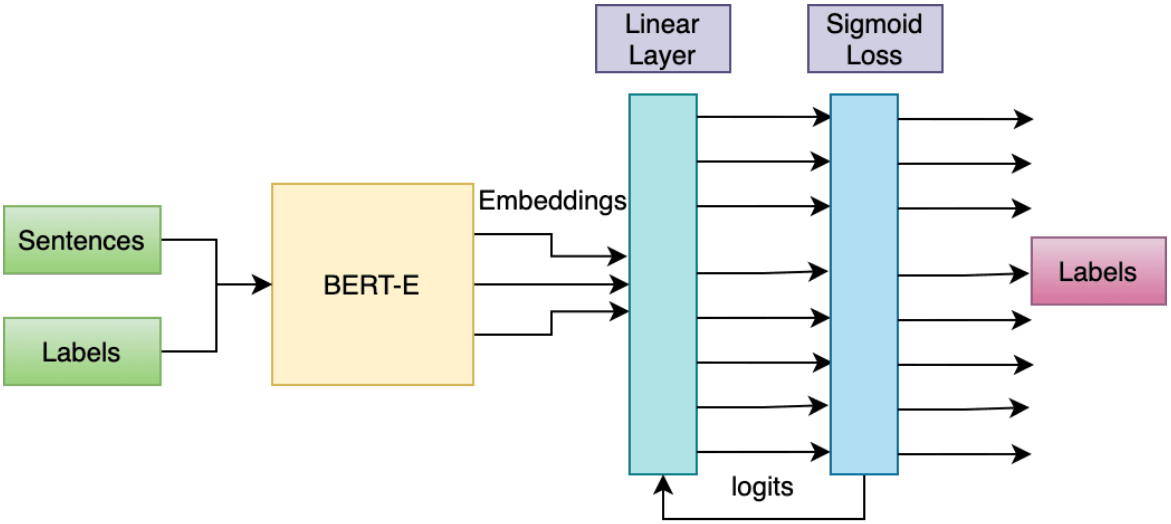


Fig: GKR Architecture

Global Change Master Directory (GCMD) Keywords are a hierarchical set of controlled Earth Science vocabularies that help ensure Earth science data, services, and variables are described in a consistent and comprehensive manner and allow for the precise searching of metadata and subsequent retrieval of data, services, and variables.

We trained a fully connected multi-class multi-label network that uses BERT-E embeddings for the text as input and sigmoid activation as the output layer.

The training was done using 7k sentences and their labels for 10 epochs using a 90-10 training and eval split. For simplicity, both models were trained to predict GCMD keywords upto the forth level of hierarchy.

It was observed that the SciBERT based model attained an accuracy of **89.99** whereas the BERT-E based model attained an accuracy of **92.18**. Here, the accuracy was measured using the following expression:

$$A = \sum \frac{preds == labels}{n(labels)}$$ where,

A = Accuracy
preds = predictions made for sentence by model
labels = true values of labels for the sentence
n(labels) = total number of labels

QUALITATIVE ANALYSIS

The following shows the comparison between BERT-E and Sci-BERT on masked language modeling task:

Sentence	Sci-BERT	BERT-E
Narrow bipolar event (NBE) is a special type of intracloud lightning {discharge} with many distinguishing characteristics	event phenomenon events , phenomena	, discharge* phenomenon flash process
More recently, ocean color satellite observations provided evidence for enhanced {phytoplankton} growth in the HNLC Northeast Pacific	ocean population phytoplankton* ice vegetation	phytoplankton* algal ice chlorophyll microbial
East African drying is linked to Indian Ocean SSTs warming {faster} than Pacific SSTs	more rather faster* greater earlier	faster* more rather earlier slower
Volcanic ash emissions and subsequent {aerosol} deposition to the surface ocean have also frequently been implicated as a source of Fe	atmospheric direct subsequent dust ash	dry atmospheric wet dust aerosol*
Clouds are an important part of the {atmospheric} radiation system	global solar total thermal diffuse	earth atmospheric* global terrestrial solar
A galactic GRB could be sufficiently distant but still provide the necessary {irradiation} so that its remnant might not have been observed	conditions information space energy data	energy information structure conditions mass

fig: SciBERT vs BERT-E for MLM

FUTURE TASKS/ACKNOWLEDGEMENTS

- Train with both NSP (Next Sentence Prediction) and MLM (Masked Language Modeling)
- Train on even bigger corpus of data
- Implement on several other downstreaming tasks such as Graph Convolutions, Satellite/Product recommendations etc
- Eventually train BERT model from scratch including the tokenizers

This research is supported by NASA Grant NNM11AA01A as part of the Interagency Implementation and Advanced Concepts Team (IMPACT) program.

ABSTRACT

Language models are fast approaching human-like understanding of natural language. They have been shown to perform equally, if not better than humans in a myriad of language tasks such as next sentence prediction, question answering, entity extraction etc. Part of the success of the models are owed to the fact that they have been trained on varied natural language text over the internet. By virtue of this, the models do not contain the semantic information present in Earth science literature. Hence, there is a lot of room for improvement when using these models for earth science specific tasks. In this work, we showcase our approach on developing Earth science specific language models. Furthermore, we justify the need for such a model by using the embeddings generated by the model to perform a domain specific downstream task that performs better than a generic model.

REFERENCES

[1] 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 Advances in Neural Information Processing Systems, pages 6000–6010

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

[3] Maskey, M.,Alemohammad, H.,Murphy, K. J., and Ramachandran, R. (2020), Advancing AI for Earth science: A data systems perspective, Eos, 101, https://doi.org/10.1029/2020EO151245. Published on 06 November 2020.