Towards an Aviation Large Language Model by Fine-tuning and Evaluating Transformers

1st David Nielsen
KBR Inc.
NASA Ames Research Center
Moffett Field, USA
david.l.nielsen@nasa.gov

2nd Stephen S. B. Clarke
Flight Research Aerospace
NASA Ames Research Center
Moffett Field, USA
stephen.s.clarke@nasa.gov

3rd Krishna M. Kalyanam
Aviation Systems Division
NASA Ames Research Center
Moffett Field, USA
krishna.m.kalyanam@nasa.gov

***Abstract*—In the aviation domain, there are many applications for machine learning and artificial intelligence tools that utilize natural language. For example, there is a desire to know the commonalities in written safety reports such as voluntary post incidents reports or aerial wildfire operations reports to better understand the risks present. Another use-case is the possibility of extracting airspace procedures and constraints currently written in documents such as Letters of Agreement. These applications can benefit from the use of state-of-the-art natural language processing techniques when adapted to the language/phraseology specific to the aviation domain. This paper evaluates the viability of adaptation of NLP tools to the aviation domain by fine-tuning transformer based models using aviation data sets.**

In 2018, a novel language model based on neural units (also called transformers) was created and became known as "Bidirectional Encoder Representations from Transformers" or BERT. This architecture combined with large amounts of English training data and innovative semi-supervised training tasks set the standard for what would later emerge as Large Language Models. The performance of these models was further improved by hyperparameter tuning and refinement of the semi-supervised training task and resulted in "Robustly Optimized BERT Pre-training Approach through hyperparameter tuning" or RoBERTa models. These pre-trained Large Language Models proved to be useful for a wide variety of natural language processing tasks such as text classification and question answering through a process called fine-tuning. The transformer architecture with pre-trained weights served as the basis with the last few layers replaced with layers fine-tuned to perform a new task e.g., a layer that provides a label for the entire input text. This process of fine-tuning can also be used to adapt the models to new domains; e.g., BioBERT started with the pre-trained BERT model and was completed by additional fine-tuning and training on biomedical documents. Transformer-based architectures can also be used to create rich representations of text called *embeddings* which can serve as the input to other machine learning models. This allows simpler algorithms such as logistic regression to use context-rich representations of the text while still remaining quick to train and evaluate.

In the world of aviation, there is a growing demand for natural language processing and understanding but the domain presents unique challenges. Due to the technical content (and specialized language) of most aviation documents, fine-tuning pre-trained Large Language Models to specific tasks has not met the benchmark on natural language processing tasks set by simpler models trained from scratch on the data. To address this deficiency, this paper evaluates the improvements from fine-tuning a Large Language Model on a large set of aviation documents using the original semi-supervised training tasks before performing specific natural language tasks. In fine-tuning, a domain-specific dataset is used on the original training task but with the pre-trained Large Language Model instead of starting from a random initialization. This approach allows the model to be adapted to the specific domain language without discarding the information gained from training on general English data.

This paper utilized two major dataset types to train and assess the RoBERTa fine-tuning performance. The first are 7,057 Letters of Agreement which are Federal Aviation Administration (FAA) documents that formalize airspace operations across the national airspace system. They contain many examples of 'aviation English' using domain specific terminology and phrasing which serves as a representative basis to perform the semi-supervised fine-tuning. The second type is the 494 document classification labels to be used for evaluation. This down-stream evaluation aims to show the performance of the fine-tuned model, better understand how much data is needed for an effective fine-tuning, and how fine-tuning can be adapted for different applications in the domain.

After semi-supervised training, evaluation begins by encoding the documents for classification using the fine-tuned RoBERTa model. Then a logistic regression classifier is trained to label the document type and compared against our ground truth labels. This currently leads to a 82.8% accuracy on 10-fold cross validation showing improvement over baseline RoBERTa which achieved 81.0%. We plan to measure the improvements on additional tasks and it is expected that these improvements will lead to more robust models that can tackle the natural language processing challenges present in aviation datasets.