VECTOR INSTITUTE | INSTITUT VECTEUR

Harnessing the Power of Natural Language Processing (NLP):

A Vector Institute Industry Collaborative Project A Technical Report

Authors (in alphabetical order)

Sedef Akinli Kocak[1*], Jimmy Ba[1,2], Elham Dolatabadi[1,2*], Garth Gibson[1], Gennady Pekhimenko[1,2], Frank Rudzicz[1,2,3,4,5]

Participating Sponsors

BMO, NVIDIA, Scotiabank, TD (Layer 6), Thomson Reuters, CIBC, Deloitte, Georgian, Intact, Manulife, PwC, Sun Life, Goldspot Discoveries, ROSS, Tealbook, Wattpad.

[1] The Vector Institute
[2] University of Toronto
[3] St. Michael's Hospital
[4] International Center for Surgical Safety
[5] Li Ka Shing Knowledge Institute

***Corresponding Authors**
sedef.kocak@vectorinstitute.ai
elham.dolatabadi@vectorsintitute.ai

Executive Summary

The Vector Institute's Recreation of Large Scale Pre-Trained Language Models project (the NLP Project) is an industry-academia collaboration that explores how state-of-the-art natural language processing (NLP) models could be applied in business and industry settings at scale.

Developing and employing NLP models in industry has become progressively more challenging as model complexity increases, data sets grow in size, and computational requirements rise. These hurdles limit the accessibility many organizations have to NLP capabilities, putting the significant benefits advanced NLP can provide out of reach. The NLP Project addressed these challenges by familiarizing industry participants with advanced NLP techniques and the workflows for developing new methods that could achieve high performance while using relatively small data sets and widely accessible computing resources.

Whereas most NLP research collaborations are designed to produce state-of-the-art models with competitively low error rates, Vector's objective was to create a collaborative and scalable learning environment that would allow several companies to gain the hands-on experience necessary to build an end-to-end NLP pipelines and scale their deployment whose primary objective is to produce business value. As such, the project involved 60 participants: 23 Vector researchers and staff with expertise in machine learning and NLP along with 37 industry technical professionals from 16 Vector sponsor companies. The participants established 11 working groups, each of which developed and performed experiments relevant to existing industry needs.

Experiments were organized in three categories:

- **Domain-specific training,** where working groups explored language models applicable to the health, finance, and legal fields for effective domain-specific training and fine-tuning.
- **Pre-training large models** where working groups trained large transformer-based language models from the ground up, with investigations of GPT-2 training and techniques to multi-node $\text{BERT}_{\text{LARGE}}$ pretraining in order to increase cost-efficiency.
- **Summarization, question answering, and machine translation** where working groups investigated NLP tasks, including health and finance-specific text summarization (extractive and abstractive), machine translation, and health-specific question answering systems in response to COVID19.

The key achievements and highlights of the project are as follows:

- ***Knowledge transfer***. A venue for a large number of industry practitioners and Vector researchers' engagement, exchange of knowledge, capacity building and skills advancement with experiential learning for a large number of practitioners working with state-of-the-art NLP models.
- ***High performance computing cluster access***. Successful development and deployment of distributed training strategies to scale up training and fine tuning of transformer based deep learning models on shared computing clusters. Vector provided computing cluster access and usage training first time to the industry participants who built their capability on using high performance computers and training large models who would not be able to do otherwise.
- ***An end-to-end NLP pipeline.*** An engagement in the entire process of building an NLP pipeline from data ingestion (e.g., web scraping) to large scale training and downstream fine-tuning.
- ***Real business impact.*** An exclusive opportunity for enhancing the capacity building of industry sponsors in the NLP domain. Insights gained in the NLP Project have informed programs and product development in participating sponsors and provided unique technology development opportunities to industry technical staff.

- ***Preparedness in response to COVID-19.*** A world class partnership with sponsor practitioners was established for the Kaggle CORD-19 dataset challenge and developed a smart system to accelerate dissemination of scientific knowledge and help the medical community in the fight against COVID-19.

Taken together, through the NLP Project, industry participants benefited by gaining experience with pre-training of large scale language models, attending expert lectures leading to effective knowledge transfer, accessing Vector's scientific computing resources, and establishing fruitful collaborations with other sponsor organizations. Notably, insights gained in the NLP Project have informed programs and product development in some participating organizations.

Table of Contents

Introduction

Vector Institute launched a multi-phase industrial-academic collaborative project in Natural Language Processing (NLP), inspired by growing interest in recent advances and breakthroughs in the field (Lan et al., 2019; Liu et al., 2019; Yang et al., 2019). This report provides an overview of the collaboration between the Vector Institute (Vector) and some of its industrial partners in that project. Vector Institute is an independent, not-for-profit corporation dedicated to advancing artificial intelligence (AI) and excelling in machine learning (ML) and deep learning (DL). Vector's vision is to drive excellence and leadership in Canada's knowledge, creation, and use of AI to foster economic growth and improve Canadians' lives.

Background:

NLP enables computers the ability to "learn" language using data, with some capabilities now approaching human-like performance. Advances in deep learning techniques (e.g., attention mechanisms, memory modules, and architecture search) have made impressive improvements in the NLP landscape (Yogatama et al., 2019). Many NLP tasks, such as question answering, machine translation, reading comprehension, sentiment analysis, and summarization, are often approached using supervised learning on task specific labeled datasets (Radford et al., 2019). However, recent breakthroughs demonstrate that models that are pre-trained on a large unlabeled corpus perform well on many NLP tasks without explicit supervision (Devlin et al., 2018; Radford et al., 2019). These models use a combination of pre-training and supervised fine-tuning where transformers (Vaswani et al., 2017) are used as the backbone of learning. During pre-training, the transformer is trained on a large corpus in an unsupervised fashion such as language modeling (predicting the next word given a context), masked language modeling (predicting a missing word in a sentence from the context), and next sentence predictions (predicting whether two sentences are consecutive sentences). The transformer is then used on various NLP tasks by adding an extra task-specific final layer for fine-tuning.

There are at least three factors that affect the performance of transformer-based pre-trained models: (1) size of corpus, (2) availability of computational resources, and (3) expressiveness of model architecture. Because of these factors, the cost and complexity of developing pre-trained models are rising quickly and limit the capability of reproducing high-performance results for those without sufficient resources.

Project Overview:

This is a joint academic-industrial collaborative project launched in Summer 2019 to explore opportunities as well as promote recent advances in the NLP domain. The project involved 60 participants: 23 Vector researchers and staff with expertise in machine learning and NLP along with 37 industry technical professionals from 16 Vector sponsor companies. The participants established 11 working groups, each of which developed and performed experiments relevant to existing industry needs.
The primary objectives of the project were:

- to foster and widen productive collaboration among academic researchers and industry practitioners on projects in the NLP domain,
- to help participants in gaining proficiency in building an end-to end NLP pipeline from data ingestion to large scale training and downstream fine-tuning, and
- to build the capacity for further advances and new lines of businesses in large scale language models in our ecosystem.

The project was conducted in three phases over 12 months (4 months for each phase); In total 11 working groups were formed to undertake different tasks and activities to achieve the project's main

objectives. During the project, weekly meetings were held as ways of communicating current updates and tasks among project members. Commonly found group activities in the weekly meetings were problem solving, decision making, prioritization, and task assignment. Weekly meetings also featured invited guest talks, tutorials on recent advances in NLP and ML from academia and industry, and reading group activities of recent related literature.

Regarding high-performance computing resources, Vector Institute provided participants with 528 GPUs, 6 GPU nodes of 8 x Titan X, and 60 GPU nodes each with 8 x T4, for development and deployment of large-scale transformer-based NLP models. For the model development and evaluation, the implementation from the pytorch transformers library by Hugging Face[2] (2019) was used.

Three main project focus areas arose which reflected current industry needs, participants' interests and expertise, and opportunities to translate academic advances into real-world application: (1) domain-specific training, (2) pre-training large NLP models, and (3) summarization, question answering, and machine translation. The remainder of this report provides a high-level overview of these focus areas and brief summaries of the working group's activities and sub-projects in each area. The participants' names are listed in the Appendix section.

[2] https://github.com/huggingface

Focus Area 1: Domain-specific Training

The goal of this focus area is to provide an effective way of domain-specific training and fine-tuning of transformer based NLP models. Working groups in this focus area created their domain-specific language models in health, finance, and law by starting domain-specific pretraining from an existing general-domain language model. Domain-specific pre-training can provide additional gains over domain-agnostic language models if text data (e.g. health, financial data) is different from the "standard" text corpus used to train BERT (Devlin et al., 2018) and other language models. This section of the report summarizes the major methodologies used/reproduced in health, finance, and legal domain-specific training.

Health Domain:

Pre-training Language Representations in the Biomedical Domain

Background/Objective:

BioBERT (Lee et al., 2020) is an extension of BERT that is further pre-trained on domain-specific biomedical corpora, including text from PubMed and PubMed Central (PMC). The primary aim of this workstream is to replicate some of the findings of the BioBERT study.

Methods:

$BERT_{BASE}$ is the baseline model. The working group started with $BERT_{BASE}$, then pre-trained it on the PubMed abstracts data ($BERT_{BASE}$ + PubMed). The PubMed corpora used for pre-training consists of paper abstract from millions of samples of biomedical text. While the original BioBERT study considers combined pre-training on PubMed[3], PMC, and PubMed + PMC together, this working group's model was pre-trained only on PubMed. This was done to check the performance using a smaller amount of data in consideration of the shared computing resources. The PubMed data was processed into a format amenable for pre-training. The raw data consisted of approximately 200 million sentences in 30 Gigabytes. The raw sentence data was batch processed into 111 chunks of ready-to-consume input data for BERT pre-training. Technically, the input for the Next Sentence Prediction (NSP) task was lower-cased sentences with a maximum length of 512 tokens and masked at the sub-token level analogous to the original $BERT_{BASE}$(uncased). The original BERT models were trained on data with maximum sequence lengths of 128 tokens for the initial 90% of the training and 512 for the remainder. Generally, a longer sequence length is preferable if the corpora tends to have longer passages but this means the amount of data and time to train also increases. In this experiment, the training was maintained at the maximum length of 512 tokens throughout. In this 4 GPU (Titan X) computing environment, one epoch of the PubMed data was passed through in approximately 895 hours. Approximately 60% of the epoch was optimized with a batch size of 28. The final 40% of data was passed through with halved resources at a batch size of 14. The full training loss of $BERT_{BASE}$(uncased) is shown in Figure 1.

Results and Discussion:

The pre-training experiment was designed to accommodate some realistic computing resource limitations. In a time-shared computing cluster, GPU resources were allocated on a limited priority basis by Slurm Workload Manager. The working group was able to successfully reproduce the results of BioBERT in part using time-shared computing resources. This work confirms that unsupervised pre-training in general

[3] https://www.ncbi.nlm.nih.gov/pubmed/

could improve the performance on fine-tuning tasks where large datasets exist. However, the effectiveness of domain-specific pre-training as a way of further improving the performance of supervised downstream tasks may not be wholly substantiated owing to a lack of consistent evidence. Moreover, to facilitate stable training and to accommodate a hardware environment where resources may be reallocated to higher priority users at any time and without warning, the following training features were implemented: (1) storing model weights and gradients at regular time intervals during training; (2) querying and automating job submission from system task scheduler; and (3) automatically restoring training from data chunk and step as GPU resources became available.

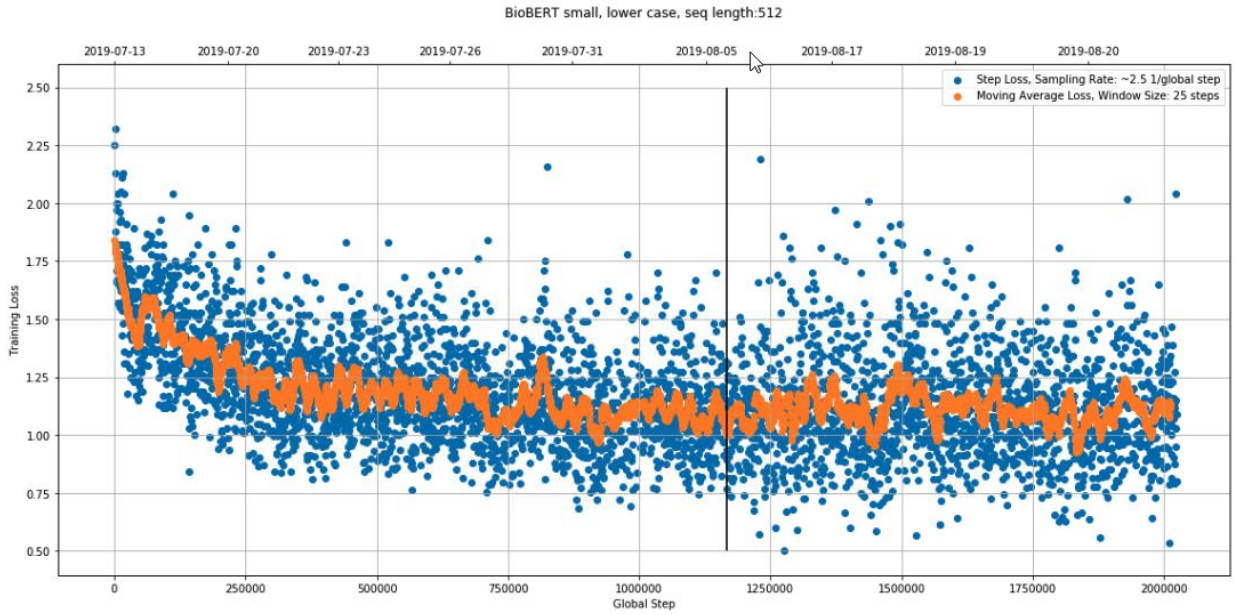

Figure 1: Training loss. The effect of chunking the data into shards can be seen in the fluctuation of the training loss. The vertical line indicates separation between batch size 28 and 14 in the training paradigm. The top (bottom) x-axis indicates calendar date (global step).

An Experimental Evaluation of Large NLP Models in the Biomedical Domain

Background/Objective:

There has been a rapid growth in NLP for the biomedical domain (Khattak et al., 2019) and state-of-the art transformer-based models are shown to achieve impressive results on various biomedical tasks through fine-tuning. The objective of this working group is to conduct an empirical investigation on fine-tuning BERT models for downstream biomedical tasks including named entity recognition (NER), relation extraction (RE), and question answering (QA). The work aims to answer whether domain-specific BERT (BioBERT) improves performance compared to BERT which has been trained on domain-agnostic corpora.

Methods:

To accomplish the objective, the working group leveraged the BioBERT model ($BERT_{BASE}$+ PubMed) built in-house, BioBERT[4] model, and the original BERT[5] model for evaluation of downstream NER, RE and QA tasks. In this work, three NER datasets including NCBIDisease (Doğan et al., 2014), JNLPBA (Kim et al., 2004), and Species-800 (Pafilis et al., 2013) were used. The GAD (Becker et al., 2004) dataset was used for the RE experiment. Regarding the QA task, the working group explored two common biomedical datasets: BioASQ (Tsatsaronis et al., 2015) and PubMedQA[6] (Jin et al., 2019). For BioASQ[7], as was suggested in the BioBERT study, all BERT models were initially fine-tuned on the SQuAD[8] (Rajpurkar et al., 2016) dataset (with intermediate evaluations), and then on the BioASQ training set before final evaluation on the BioASQ test sets. Three BioASQ tasks including 4b, 5b, and 6b were used in the experiment. For the PubmedQA, all BERT models were fine-tuned on PQA-L which is 1k expert-annotated generated QA instances.

Results and Discussion:

For the NER experiment, the working group found that the F1-scores for the vanilla $BERT_{BASE}$ were higher on average than the ones fine-tuned with the in-house $BERT_{BASE}$+ PubMed. For RE, the group's $BERT_{BASE}$+PubMed outperformed other BERT models, but the differences between models was not significant. Regarding the BioASQ and PubMedQA, the results indicated that pre-training on biomedical domain corpus improves performance on the downstream BioASQ QA task. However, the improvement is not so large as to be entirely convincing and carefully fine-tuned BERT models can perform comparably to BioBERT. Overall, the group's exploration confirms that unsupervised pre-training in general could improve the performance on fine-tuning tasks. However, the effectiveness of domain-specific pre-training as a way of further improving the performance of supervised downstream tasks does not significantly outperform the effectiveness of domain-agnostic BERT models considering the high cost of domain-specific pre-training which makes it challenging for most researchers and NLP developers. In the biomedical domain, however, this conclusion may not be wholly substantiated owing to a lack of consistent evidence, particularly in downstream NER and QA tasks.

Finance Domain:

Sentiment Classification and Extractive Summarization on Financial Text Using BERT

Background/Objective:

Relatively little evidence about performance of transformer-based language models in the finance domain exists. The goal of this work was to fill this gap by effectively fine-tuning BERT on finance-specific data so that NLP researchers in the finance domain could potentially benefit from it.

Methods:

In this workstream FinanceBERT-SUM is proposed where the working group performed domain-specific pre-training of BERT (Devlin et al., 2018) using examples portraying financial elements. First, the group

[4] https://github.com/dmis-lab/biobert
[5] https://github.com/google-research/bert
[6] https://pubmedqa.github.io/
[7] http://bioasq.org/
[8] https://rajpurkar.github.io/SQuAD-explorer/

scraped large amounts of financial text from the web. Then, the group provided a semiautomated approach using latent dirichlet allocation (LDA) to extract stories explaining financial topics and create a financial version of the CNN/Daily Mail (Hermann et al., 2015) dataset. One advantage of using BERT is that any task-specific layer can be plugged in at the end as the final layer without designing any substantial architecture modifications. Leveraging from this the working group performed two different experiments. In the first one, the group used a BERTSUM (Yang Liu, 2019) model pre-trained on general domain corpora and further tuned it on the finance data crawled from web pages. Later, the group fine-tuned it further on the training set of the financial version of CNN/Daily Mail dataset. Similar to (Yang Liu, 2019), the group used three different types of classifiers in the final layer: (1) a simple classifier with just one fully connected layer; (2) a transformer layer with two intermediate layers followed by a fully connected layer; and (3) a recurrent neural network followed by a fully connected layer. In the second experiment, the group took the vanilla BERT (Devlin et al., 2018) model trained on general domain corpora and fine-tuned it only on the training set of FiQA (Maia et al., 2018) and Financial Phrase Bank (Malo et al., 2014) dataset. The group skipped the second stage of fine-tuning using the crawled data as the model became overfitted.

Results and Discussion:

In this work, the group proposed an effective domain-specific pre-training strategy for large-scale language models in the financial domain and proposed FinanceBERT (-CLS and -SUM), having specialization on financial sentiment analysis (FinanceBERT-CLS) and financial summarization tasks (FinanceBERT-SUM). The group developed a semi-automated strategy of separating finance-related data from a collection of general-purpose data, a strategy that can be applied to any other domain. All the fold performance results were similar to the literature (Araci, 2019) and this behavior was very consistent over all the performance metrics. Experiments showed the validity and quality of the finance data that the group created and the effectiveness of performing domain-specific pre-training.

Legal Domain:

Investigation of Transformer-based Model in Legal Texts

Background/Objective:

Recent advancements in pre-training have enabled language models to perceive the semantic and syntactic essence of a language structure well. However, exploiting them in a real-world domain-specific scenario still requires some practical considerations to be taken into account. These include such token distribution shifts, inference time, memory, and their simultaneous proficiency in multiple tasks. The working group's objective was to investigate various approaches to customize a transformer-based language model to a legal domain and to compare performance in terms of downstream tasks and time/memory considerations.

Methods:

The working group used a publicly-available corpus of 9,000 U.S. legal agreements[9] as the domain-specific text data. The group investigated the impact of two main factors on training of language models: tokenization and weights initialization. In one experiment, the group used SentencePiece on the legal corpus to generate the same number of cased tokens as $BERT_{BASE}$(cased)'s general tokens. The group referred to these domain-specific tokens as 'legal tokens'. Only 36% of tokens are common between legal tokens and general BERT tokens. The group also used a hybrid version of tokenization in

[9]https://www.sec.gov/edgar.shtml

which it only added the 500 most frequent tokens in the legal corpus that do not exist as independent tokens in the general BERT. For general and hybrid tokenization (500 legal tokens + BERT tokens) approaches, the group started the training both from the general-domain model weights published with the original papers (i.e., pre-trained initial weights) and from scratch (i.e., random initial weights). Therefore, the group compared six variations of the BERT model:

- Base: $BERT_{BASE}$(cased) without any customization for the legal domain;
- GR: General Tokens with Random initial weights (we pre-train from scratch);
- GP: General Tokens with Pre-trained initial weights;
- LR: Legal Tokens with Random initial weights;
- HR: Hybrid Tokens with Random initial weights;
- HP: Hybrid Tokens with Pre-trained initial weights.

In addition to BERT, the group trained DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2019). ALBERT and RoBERTa were only pre-trained on the corpus by using the general pre-trained weights. Four downstream tasks were defined: Passage Retrieval (PR); Named Entity Recognition (NER); Text Matching (TM); and Sentiment Analysis (SA).

Results and Discussion:

The group observed the training loss moving average through 10 epochs of training the BERT model training on a 4 GPU with batch size of 32 and ADAM optimizer. The training loss saturated much faster when starting from the pre-trained weights. Also, when comparing GR, HR, and LR models, the group noted that adding more domain-specific tokens delays the saturation in loss. For all of the downstream tasks except the NER task, the group added a fully connected layer on top of the [CLS] embeddings (or its equivalent in the model). In the NER task, the group used a fully connected layer applied to each token embedding to determine the labels in the experiments. In all of the fine-tuning experiments, the group trained all of the weights (language model + fully connected). For all four tasks, the group split the datasets into training, validation, and test sets in proportions of 80%, 10% and 10% respectively. In each task, the group used early stopping based on the validation loss for all models.

In PR, TM, and SA, pre-training the general BERT language model on the corpus improved the performance results on average 8.5% over the base version. In these three tasks, the highest performance of DistilBERT, RoBERTa, and ALBERT could achieve 95%, 97%, and 95% respectively. These models also seemed to benefit from domain-specific language model customization. For the NER task, the base language models seem to perform better in general. Moreover, by comparing GR and LR versions of BERT and DistilBERT, the group noted that using legal tokens marginally improves the performance compared to using the default general-domain when starting from scratch. However, it still does not beat the impact of using pre-trained weights considering the size of the corpus and the amount of language-model training performed (10 epochs). Even extending the pre-trained model with only some legal tokens degrades the performance for most of the tasks (Shaghaghian et al., 2020).

Details are presented in: Shaghaghian, Shohreh, Luna Feng, Borna Jafarpour and Nicolai Pogrebnyakov, "Customizing Contextualized Language Models for Legal Document Reviews", the Fourth Annual Workshop on Applications of Artificial Intelligence in the Legal Industry, IEEE Big Data Conference. 2020.

Focus Area 2: Pre-Training Large Models

The working groups in this focus area explored design and implementation of transformer-based language models pre-training using Vector's high performance computing cluster aiming to make

cutting-edge NLP models more accessible. Pre-training of language models is a heavy computational lift that often requires days of processing on costly infrastructure exceeding many organizations' and academic institutes' resources. One group focused on providing algorithmic and software optimizations solutions to address this limitation. The other group investigated system engineering challenges associated with training a large language model in an academic institute such as GPT-2.

Multi-node BERT-Pretraining: Cost-efficient Approach

Background/Objective:

The BERT language model has significantly improved the state-of-the-art performance of many downstream NLP tasks such as language understanding and question answering. However, training BERT is computationally intensive due to its high model complexity (110M parameters in $BERT_{BASE}$ and 340M parameters in $BERT_{LARGE}$) and large amount of training data. Programs often require an advanced hardware setup to train these models within reasonable time. Transformer-based models provide opportunities to train BERT on multiple GPU/TPU devices, which linearly reduces the training time according to the amount of hardware resources available. However, machines that are capable of servicing such levels of parallelism are usually very expensive and overly dedicated for computationally intensive workloads. The objective of this working group was to address this limitation and complete $BERT_{LARGE}$ pre-training on a cluster of widely available GPUs through careful optimizations and massively parallel workload organization.

Methods:

Like the original work on BERT by Devlin et al. (2018), this working group used the Wikipedia Corpus[10] and BookCorpus[11] datasets for pre-training. After extracting plain English text from those two public datasets, the group processed the sentences. First, the group tokenized the raw text, then masked out 15% of the words in the input sentences in order for the model to learn the relationship within the sentence. Finally, the group split and shuffled adjacent sentences with 50% probability for the model to do next sentence prediction.

The group used 32 nodes connected by a 10Gb/s network with a total 256GPU (NVIDIA T4) with 1024 total CPU (32 CPU per node (Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz)). A detailed account of the system configuration can be found in Lin et al. (2020). The group trained $BERT_{LARGE}$ on the combined Wikipedia Corpus and BooksCorpus dataset. The group then followed a two-phase training schema when performing pre-training, using sequence length of 128 tokens to train the first 36 epochs and a sequence length of 512 tokens to train the last 4 epochs. To improve the single device throughput and fine-tune the training workload with respect to the distributed system topology, the group performed the following optimizations on $BERT_{LARGE}$ pre-training:

1. data sharding before training based on the number of devices to address the I/O bottleneck;
2. automated mixed precision to reduce computing and memory requirements;
3. kernel fusion to reduce the overhead of kernel execution on hardware;
4. data parallelism to scale to multiple GPUs in a multi-node context; and
5. gradient accumulation to reduce the communication demands that arise in distributed training.

For fine-tuning, the group picked the question-answering task trained on the Stanford Question Answering Dataset (SQuAD1.0) (Rajpurkar et al. 2016). This dataset contains 100,000 question-answer pairs from Wikipedia, in which a question and a passage are provided as the training data and the corresponding answer is provided as the training label.

[10] https://github.com/attardi/wikiextractor
[11] https://github.com/soskek/bookcorpus

Results and Discussion:

The group presented approaches on how to improve single device training throughput, distribute the training workload over multiple nodes and GPUs, and overcome the communication bottleneck introduced by the large data exchanges over the network. The group demonstrated the ability to perform pre-training on BERT within a reasonable time budget in an academic setting, and with a much less expensive and more generic hardware resource requirement than typically assumed required in an industrial setting. The group was able to achieve a 70% weak scaling efficiency and complete BERT training in 12 days - a reasonable time frame in an academic setting. The model achieved 81% to 83% F1 scores depending on the loaded pre-trained checkpoints. This represents a discrepancy of approximately 9% to 10% when compared with the Devlin et al. (2018) result of 90.9% and NVIDIA's result of 90% to 91%[12].

Details are given in: Lin, Jiahuang, Xin Li, and Gennady Pekhimenko. "Multi-node Bert-pretraining: Cost-efficient Approach." arXiv preprint arXiv:2008.00177 (2020).

Investigation of GPT-2 Training on Vector Computer Cluster

Background/Objective:

This work aimed to provide an overview and discussion of the challenges of training a GPT-2 model from a system perspective. GPT-2 (the successor to GPT) (Radford et al., 2019) is a large transformer-based language model with 1.5 billion parameters trained for predicting the next word in a sentence, given all of the previous words used within some text. GPT-2 is a direct scale-up of GPT, with more than 10 times the parameters and trained on more than 10 times the amount of data.[13]

Methods:

The dataset used in this study was OpenWebText[14], a public replication of WebText corpus used to train GPT-2 (Radford et al., 2019). The format and contents follow WebText. The format and contents follow WebText. The text extracted from articles linked in Reddit posts that received at least three Reddit 'karma'. Due to the memory constraints, the working group split the data into 128 shards, with each containing approximately 100 million tokens. Data preprocessing followed Byte Pair Encoding (Sennrich et al., 2015). The model for this study was GPT-2, implemented in Tensor2Tensor[15] to investigate end-to-end training time. Most of the experiments were conducted with GPT-2 small, and were designed to be as close as possible to the original GPT-2 small model. The experiments were conducted on two nodes with total 8 GPUs (NVIDIA P100) with a multi-step optimization setting to simulate training with a batch size of 524,288. The step size for single node and two-node experiments is 64 and 32 respectively. The group experimented with time spent on the forward pass and back propagation in each iteration on a single node and on two nodes. The aim was to test out the total time spent on data loading and preprocessing on a single node and distributed nodes, and the total time spent not performing the training computation during each iteration.

[12] https://github.com/NVIDIA/DeepLearningExamples
[13] https://openai.com/blog/better-language-models/
[14] https://github.com/jcpeterson/openwebtext
[15] https://ai.googleblog.com/2017/06/accelerating-deep-learning-research.html

Results and Discussion:

The group expects that loading and preprocessing the training data, training the embedding layers or copying the model parameters to different GPUs and aggregating the gradients may be contributors to the total time spent that is not on performing the training computation during each iteration. The group observed that in the data, preparation, loading and preprocessing, a batch of 2048 tokens per GPU takes roughly 3 to 4 seconds for every 100 iterations. This may suggest that data preparation does not contribute much to the training time. The time for distributed training of GPT-2 small suggests that the forward and backward pass take roughly 30 and 32 seconds respectively, compared to 15 and 20 seconds on a single node. The time spent before performing the training computations increased significantly from 24.4 seconds to 149.6 seconds. One possible explanation for this is that since there are 2 nodes allocated for training, the number of training examples requiring preprocessing and the number of GPUs requiring feeding doubles. The group also observed communication bandwidth was bottleneck for training, even within a single node.

Investigations on Robust Representation and Fixup Initialization for RoBERTa

Background/Objective:

The goal of this group was to test whether academic insights related to pre-training language models could be translated into novel research valuable to industry. This working group created an experimentation cycle in which the main research question was first developed using Vector insights and knowledge, data preparation was conducted, and basic experimentation ensued. Promising results would lead to a continuation of this research project internally. Inspired by Phan et al. (2019), the first aspect of the group work corresponded to robust representation of scientific names. More specifically, the group focused on the unsupervised task of word vector representation, similar to word2vec. This focuses on the conceptual meanings of the scientific terms, a challenge due to the different name variation in scientific documents. The second aspect aimed to remove the dependence on warmup phase and LayerNorm (Ba et al., 2016) on RoBERTa (Liu et al., 2019). The working group followed the idea of Huang et al. (2020) in which this is achieved by correctly modifying the initialization scheme.

Methods:

At a high level, the strategy follows cycles consisting of: ideation phase, data preparation, baseline, experiments, and results evaluation. For the reference work (Li et al., 2004) baseline, the working group used the NCBI-Disease[16], BC5CDR-Disease[17], and BC5CDR-Chemical[18] datasets. Data is readily available and preprocessing consists of standard text cleaning techniques. The baseline architecture consists of a bidirectional-based architecture (BiLSTM) in order to extract contextual representations. The working group found several inconsistencies and gaps in its attempt to reproduce the reference work. Specifically, the published code is incomplete, and it can only be used for part of the inference process, and there is no clear evaluation. The group attempted to rebuild the code, but the results obtained were no closer to those reported. Nevertheless, the group replaced the BiLSTM portion with a transformer one, and found that it could match the results but was active with no gain. The group used the WikiText-103[19] dataset, preprocessed it via GPT-2 BPE, as the dataset for th e RoBERTa experiments. This is a small dataset, and the goal was to show removing the warmup phase and LayerNorm is possible for the initial part of training. The group trained the models up to 5000 steps. As a baseline the group used the

[16] https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
[17] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/
[18] https://www.ncbi.nlm.nih.gov/research/bionlp/Data/
[19] https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/

instructions provided by Fairseq[20]. The group trained the same model without warmup to show the evidence that is needed. The group then modified the RoBERTa architecture and removed both LayerNorm and warmup, and trained on the same dataset.

Results and Discussion:

The lack of a proper benchmark along with the poor gains on the change of architecture helped the group determine that this was not a viable path to pursue. The group quickly moved on to the second aspect of the work. The results of the training for RoBERTa can be found in Figure 2. The orange represents the group's baseline. The red one corresponds to training without warmup and the gray one is the modified version of Roberta using T-Fixup. Similarly to the results in Huang et al. (2020) the group noted that training is possible, and expected that the modified model would "catch up" and overtake the benchmark in later updates, a phenomena also observed in loc. cit. Overall, the group considers this a success and the research has been taken to the next stage.

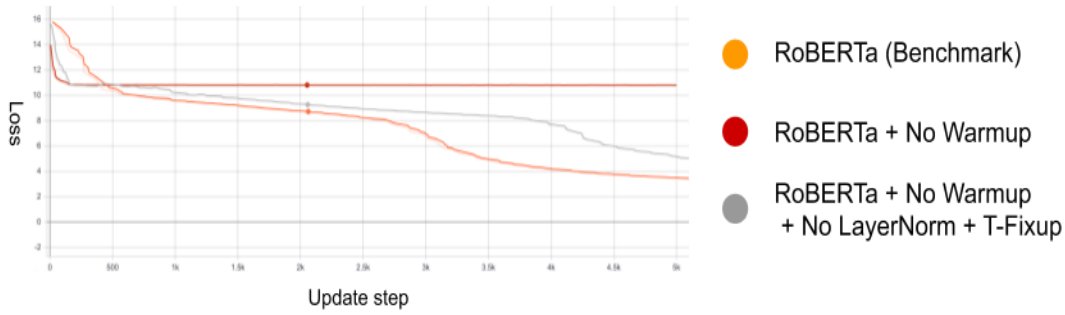

Figure 2: The results of training for RoBERTa

Focus Area 3: Summarization, Question Answering, and Machine Translation

The working groups in this focus area developed frameworks for text summarization (extractive and abstractive) in finance and medical domains, machine translation, and medical question answering systems. A special interest group (SIG-Kaggle-COVID19) formed under this focus area to create a question answering pipeline in response to COVID-19 Open Research Dataset Challenge (CORD-19)[21].

Domain-Specific Text Summarization and Dataset Generation

Background/Objective:

Automated summarization is an area of significant opportunity across a number of industries, with particularly great potential in finance and healthcare. While extensive progress has been achieved in methods for performing extractive and abstractive summarization, there is still a relative scarcity of reliable and sufficiently large datasets for model training and evaluation. In specialized domains such as health and finance, it is exceptionally difficult to find openly available, reliable, and sufficiently large annotated datasets for fine-tuning pretrained models. Recognizing the gap in specialized domain-specific datasets, the objective of this working group is to leverage the unique community structure of Reddit by

[20] https://github.com/pytorch/fairseq/tree/master/examples/roberta
[21] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

creating corpora specific to particular topics and communities represented by a subreddit or group of subreddits.

Methods:

This working group built upon the work of Völske et al. (2017), enhancing it in a number of ways to increase the precision and recall of the summaries being extracted. For data preparation, the group first cleaned and tokenized extracted post and summary pairs. The group used BART (Lewis et al., 2019), a transformer-based denoising autoencoder for pre-training sequence-to-sequence models, as an abstractive summary generator and adapted the original work for sequence generation through an autoregressive approach. Two variations of experiments were performed to benchmark domain-specific performance: in the first, the group trained the model on the entire dataset of Reddit posts and evaluated only on finance-related posts. In the second, the group used only finance-related posts for both training and evaluation. During generation, the group utilized teacher forcing, where true labels from the previous time step are used as input to generate the next summary word. For the BART architecture, the group set the number of beams to be between 1 and 10, gradient accumulation to be in 128/256/512, dropout probability of 0.1, number of encoder and decoder attention heads of 12, a repetition penalty of 2.5, a length penalty of 1, maximum input sequence length of 512 tokens, maximum output sequence length of 120 tokens, a learning rate of $1e^{-5}$ and an L2-lamda of 0.01. During training, the group unfreeze all of the 12 encoder and decoder layers. A cluster with one NVIDIA Titan XP GPU, 8 vCPUs and 12 GiB of host memory was used to train the models. During training, the group set the maximum number of attempts for validation loss increase to be 5 as the stopping criterion. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, specifically ROUGE-1 and ROUGE-2, and were used to evaluate the quality of the abstractive summaries against human-produced reference summaries.

Results and Discussion:

The main contribution of this work is the ability to scalably extract high-quality post summaries from domain-specific subreddits in a scalable way. Compared to the current baseline[22] this working group was able to significantly improve the volume and veracity of extracted TLDRs' post summaries created by Reddit users and referred to as "too long/didn't read". The group demonstrated the potential application of their data extraction and pre-training pipeline in the fields of finance and healthcare. The models can be used as standalone text summarization models, or further improved by augmenting training data with additional public or proprietary data sources. More importantly, the group demonstrated a clear positive trend in the quality of text summarization for a large language model such as BART when fine-tuned using specialized data, even in cases where the number of labeled examples available is relatively small. The group also showed the relative extent to which additional specialization can improve the quality of summarization, both quantitatively using ROUGE scores, and qualitatively using a number of illustrative examples. The group hopes these data points can be used by industry practitioners to help with the cost-benefit analysis for selecting the right level of fine-tuning, from using large out-of-the-box training models to obtaining potentially costly high-quality specialized data.

Text Summarization of Biomedical Data

Background/Objective:

While there have been significant advances in developing text summarization methods in recent years, they have largely been limited to extractive approaches in the biomedical domain. The objective of this work was to investigate fine-tuning BERT models for abstractive summarization and extractive summarization tasks in the biomedical domain. An extractive summarization task creates a summary by

[22]https://github.com/webis-de/webis-tldr-17-corpus

identifying (and subsequently concatenating) the most important sentences in a document (Liu and Lapata, 2019), while an abstractive summarization task regenerates a piece of text to provide an overview of the original document information (Du et al., 2020). In the latter, the regenerated sentence(s) may contain words that do not appear in the original document (Lin and Ng, 2019).

Methods:

The approach for this study involved implementing and analyzing abstractive and extractive text summarization for general language as well as biomedical domain-specific text. To accomplish the objective, the working group used the original BERT[23] model for evaluation of abstractive and extractive summarization. In this work BioASQ 7b[24] was used to compare the results of Liu and Lapata (2019) using the CNN/DailyMail dataset. For preparation of the BioASQ 7b, the group first extracted content of research papers from links in the dataset, and combined them with summaries provided in the dataset to form raw data. Then, the group tokenized and splitted the data into train, test and validation sets. $BERT_{BASE}$ (uncased) version of BERT was trained from scratch on BioASQ 7b to implement BERTSUM (Liu 2019). BERTSUM is an extension of BERT for summarization. Special tokens [CLS] were inserted in BERT to learn sentence representations and interval segmentation embeddings used to distinguish multiple sentences. BERTSUM extends BERT by inserting multiple [CLS] symbols to learn sentence representations and using interval segmentation embeddings to distinguish multiple sentences. BERTSUMEXT (Liu and Lapata 2019) captures document-level features to extract sentences which are constructed over the BERT encoder by stacking several inter-sentence transformer layers, and BERTSUMABS (Liu and Lapata, 2019) based on an encoder-decoder architecture in which the encoder is pre-trained and the decoder is a randomly initialized transformer which is trained from scratch. All models were trained for 40,000 steps (encoder warmup steps with 20,000 and decoder warmup steps with 10,000) with an encoder learning rate of 0.001 and decoder learning rate of 0.1 on 4 GPUs (NVIDIA TItan X).

Results and Discussion:

Upon comparing the results with CNN/DailyMail found in Liu and Lapata (2019), all summarization ROUGE F1 scores for bigrams were better in case of BioASQ 7b, whereas unigram and longest common sequence overlap showed better results on CNN/DailyMail. Among BERTSUM variants, BERTSUMEXT performed better on CNN/DailyMail, whereas BERTSUMABS performed better in BioASQ 7b. The group would expect models with extractive to perform better on datasets with (mostly) extractive summaries, and abstractive models to perform more rewrite operations on datasets with abstractive summaries. CNN/DailyMail is somewhat extractive, while BioASQ is abstractive since regenerated sentence(s) may contain words that may not appear in the original document. The group observed higher ROUGE-2 scores for BioASQ 7b on the BERT variant models than CNN/DailyMail. This result is likely due to differences in distribution of source text and summary lengths. Lower ROUGE-1 and ROUGE-L scores in the BioASQ may represent insufficient biomedical text training. In conclusion, the performance achievements are not very prominent on all the evaluation metrics of biomedical dataset. Only ROUGE-1 of BERTSUMEXTABS results of BioASQ 7b are somewhat comparable with CNN/DailyMail. The group focused on only one biomedical dataset for summarization. Du et al. (2020) shows prior domain knowledge effectively improves the performance of summarization tasks in the biomedical domain. For future work, BioBERT may be investigated for extractive and abstractive summarization.

[23] https://github.com/google-research/bert
[24] http://bioasq.org/participate/challenges_year_7

Exploration of MASS Multi-node Unsupervised Machine Translation

Background/Objective:

Inspired by the success of BERT, Masked Sequence to Sequence Pre-training (MASS) for encoder-decoder based language generation was proposed by Song et al. (2019). MASS adopts the encoder-decoder framework to reconstruct a sentence fragment given the remaining part of the sentence: its encoder takes a sentence with randomly masked fragment (several consecutive tokens) as input, and its decoder tries to predict this masked fragment. The objective of this work is to reproduce results related to unsupervised machine translation via monolingual pre-training published by Song et al. (2019).

Methods:

The main library used for this work is MASS library[25]. The working group used all of the monolingual data from News Crawl datasets for three languages German, English, and French. Data was sourced from Statistical Machine Translation[26]. Pre-training consisted of a mix of masked and causal language model prediction, performed separately within each monolingual data but while keeping a language embedding placeholder, as outlined in the reference work. The group fine-tuned the pre-trained model as part of the unsupervised pre-training steps and results were obtained via back-translation on synthetically generated data. Performance of experiment was measured on test annotated data. Experiments run mostly on 4 GPUs (T4) per language pair (eng-fr; de-en). The group used the ADAM (Kingma and Ba, 2014) optimizer with 1e-4 learning rate for both pairs, and FP16 hybrid precision with automatic mixed precision from NVIDIA AMP[27] (instead of the original FP32 as in the original work). The group used the BLEU score for comparison of its results with the reference work.

Results and Discussion:

The group ran unsupervised pre-training over two language pairs: English-French, and German-English. Monolingual training data amounted to approximately ~250 million total sentences for the English-French language pair, and 50 million sentences for the German-English language pair. The group started at first with the English-French pair, but the BLEU score results didn't seem to scale as fast as desired; training occurred for ~10 weeks on 4 GPUs. The group switched to the German-English pair and ran it over the course of ~12 weeks (approximately 800,000 iterations for German-English). The group used NVIDIA AMP since running on T4 GPUs would be 10 to 20 times slower than V100 GPUs, making the experiment very difficult to reproduce. The group's best results are as follows (BLEU scores on test sets): English-French: 9.3, French-Eglish: 10.78 (MASS reference work: English-French: 37.50, French-English: 34.90); English-German: 3.44, German-English: 7.86 (MASS reference work: English-German: 28.30, German-English: 35.20). The group observed a significant difference with respect to the best published results of the reference work. It is not clear for how long the model has been trained in the reference work. One difference in the experimental setting is that in the reference work 8 V100 GPUs were used, while the group had access mostly to 4 GPUs (NVIDIA T4) which are known to be less performant of V100, with approximately at least 50% less compute power each.

[25] https://github.com/microsoft/MASS

[26] http://statmt.org/

[27] https://developer.nvidia.com/automatic-mixed-precision

Question Answering Systems in Responding to COVID-19 Open Research Dataset Challenge

Background/Objective:

The development of question answering (QA) systems is necessary for rapidly emerging domains, such as the ongoing coronavirus disease of 2019 (COVID-19) pandemic. This is particularly true when no suitable domain-specific resources are likely available at the starting. To respond to the needs of medical experts who need to quickly and accurately receive answers to their scientific questions related to coronaviruses, researchers could develop QA systems based on articles related to COVID-19. Thus, Kaggle opened a competition named the COVID-19 Open Research Dataset (CORD-19)[28] and proposed the CORD-19 dataset that encompasses 120,000 articles about coronaviruses and other diseases. The competition offered more than 10 tasks to cover some fundamental questions related to COVID 19 and provided the chance for the ML community to develop QA systems and employ them on the CORD-19 dataset. A special interest group (SIG-Kaggle-COVID19) was established with the objective of developing question answering approaches that can help the medical community develop answers to high priority scientific questions. In order to accomplish this objective, the group developed several strategies consisting of transformer and rule-based methods and submitted four different approaches to the Kaggle-COVID-19 Open Research Dataset Challenge. Details of the group work are below.

More Question than Answers: A Rapid Response to COVID-19 Question-Answering

In this work, the working group demonstrated the development of a simple question answering system combining traditional ML tools such as LDA and k-means clustering with information retrieval and elastic search. The group showed that they could extract reasonable answers for questions related to the treatment of COVID-19. The group first focused specifically on finding answers to this question: "What is the efficacy of novel therapeutics being tested currently?"[29] for the Kagglechallenge. Then, the group advanced their work and developed a method to create QA pairs from a large semi-structured dataset through the use of transformer and rule-based models. The group proposed a means of engaging subject matter experts (SMEs) from the medical community for annotating the QA pairs through the usage of a web application and developed a web application to serve the purpose of providing an efficient user interface for annotating the QA pairs generated by the designed system. Finally, the group leveraged active learning strategies to significantly reduce the required annotation effort from the SMEs. The group empirically compared the performance of the two XGBoost-based models with DistilBERT (base-cased) (Sanh et al. 2019)

when trained on the entire CORD-19 dataset. The group observed that the DistillBERT model outperformed the XGBoost-based models substantially due to its more advanced architecture in which the embedding layers of the network are also updated during the training whereas the sentence embeddings used in the XGBoost-based models are static (Bhambhoria et al. 2020). *Details are given in: Bhambhoria, R., Feng, L., Sepehr, D., Chen, J., Cowling, C., Kocak, S., & Dolatabadi, E. (2020, November). A Smart System to Generate and Validate Question Answer Pairs for COVI D-19 Literature. In Proceedings of the First Workshop on Scholarly Document Processing (pp. 20-30).*

SBERT+BERT for Cord-19 Data

In this work, the group demonstrated the development of a question answering framework combination of

[28] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge
[29] Details are given in the Kaggle-COVID-19 Open Research Dataset Challenge submission: https://www.kaggle.com/rohanvb/cord-19-task4

Sentence-BERT (Reimers and Gurevych, 2019) and BERT models. The group showed that they could extract reasonable answers for questions related to the treatment of COVID-19. The group focusedspecifically on finding answers to this question: “What is the efficacy of novel therapeutics being tested currently?” for the Kaggle challenge[30].

Extract Excerpt from Abstract Using LDA and Fine-Tuned ALBERT

In this work, the group developed a framework utilizing topic modeling to extract the relevant articles to COVID-19. Then, they used a pre-trained ALBERT (Lan et al., 2019) model to get an excerpt from each article obtained from the previous methods that get answer questions related to COVID-19. The group first utilized the SQuAD[31] dataset so that the model targets explicitly the QA task. After pretraining on SQuAD v1.1, the group fine-tuned their model to the medical domain using BioASQ factoid QA pairs. The main advantage of this framework is that it can be applied to any queries related to COVID-19, as there is no additional tuning required. The group showed that they could extract reasonable answers for questions related to the treatment of COVID-19. The group focused specifically on finding answers to this question: “What is the efficacy of novel therapeutics being tested currently?” for the Kaggle challenge[32].

Non-Pharmaceutical Intervention (NPI) Discovery with Topic Modeling and NPI-Context: Intervention Events to Inform Search

In this work, the group considered the task of discovering categories of non-pharmaceutical interventions during the evolving COVID-19 pandemic. The group explored topic modeling LDA- and HDP-based methods on two corpora with national and international scope. The group demonstrated that these methods discover existing categories when compared with human intervention labels while reduced human effort needed. As the results have verified, combining coherence, topic similarity, and coverage can provide intervention category suggestions for experts to explore and find new NPIs leveraging national and international data[33] (Smith et al., 2020).

Details are given in: Smith, J., Ghotbi, B., Yi, S., & Parsapoor, M. (2020). ArXiv preprint arXiv:2009.13602.

Limitations and Best Practices

Limitations

Technical:

Some technical challenges that were encountered included:

- **Datasets**: Size, appropriateness, and lack of dataset in some of the domains (e.g., finance and legal) were the main limitations for this project. In order to address some of the data related challenges, different techniques such as embeddings and projection size were used. Some working groups also created their own datasets.
- **Model sensitivity and hyperparameters:** Sensitivity and hyperparameters of such large scale models would affect the reproducibility of a model. To improve performance, the working groups

[30] Details are given in the Kaggle-COVID-19 Open Research Dataset Challenge submission: https://www.kaggle.com/hngai29/sbert-bert-for-cord-19-data
[31] https://rajpurkar.github.io/SQuAD-explorer/
[32] Details are given in the Kaggle-COVID-19 Open Research Dataset Challenge submission: https://www.kaggle.com/parkyoona/lda-albert-for-cord-19-data
[33] Details are given in the Kaggle-COVID-19 Open Research Dataset Challenge submission: https://www.kaggle.com/jajsmith/npi-context-intervention-events-to-inform-search

have tuned different hyperparameters such as batch size, step size, learning rate scheduling, and iterations.

- **Computational resources:** Due to the shared computing resources at Vector, the participants were given limited access to the cluster computing. Although computer resource constraints were able to be managed by increasing capacity, it still meant that training would take much longer in real time than in actual compute time.

Project:

This collaboration and the results of its experiments proved to be a unique opportunity for industry and researcher participants, but it also presented some organizational challenges. Some project challenges that were encountered included:

- **Self-organizing group development and work distribution:** There are many benefits of self-organization such as flexibility, fast decision-making, and synergy, but there are also some challenges. At the beginning of the project, the participants were invited to spontaneously organize working groups around common project ideas. This was challenging since the participants were new to each other and coming from different organizational cultures. This challenge was overcome by the group members settling on some common goals and committing to the team effort.
- **Time commitment:** Industry participants of a given group very often have limited time to contribute and availability often fluctuates unexpectedly. The working groups organized themselves in a flexible and adaptable way to cope with change and the diverse priorities of the participants.
- **Participant turnover and knowledge localization**: Due to the time constraints of the participants, frequent turnover occurred. To mitigate the negative impact of turnover, robust project onboarding content was developed and best practices along with adequate documentation were maintained to ensure knowledge would not be localized.
- **Meeting expectations:** Weekly meetings were used for topic-related seminars and for working groups to come together and work towards achieving their goals. These meetings also accommodated larger group discussion sessions which could be overwhelming for participants. In order to make meetings more efficient, a proper agile standup format was set out with follow-up discussions afterwards.

Best Practices

Technical:

- Domain-specific pre-training is very effective even with a smaller training set and fine-tuning.
- Fully trained models can benefit many downstream tasks such as sentiment classification, summarization, question answering and domain-specific search.
- Large-scale language models seem to benefit in their downstream tasks from further pre-training on a relevant corpus.
- BERT and all its variations comfortably outperformed all of the experimented baselines, all of which had RNN-based architecture. Even though these transformer-based language models are expensive and time-consuming to train, they seem to be worth the investment.
- Successful development and deployment of distributed training strategies to scale up training and fine tuning of transformer based deep learning models on shared computing clusters.

Project:

- Identify 'quick wins' to generate early success and momentum.

- This is a new way of helping companies benefit from new advancements.
- Monitor progress and impact regularly. The roles and responsibilities were defined and the operational plan for the collaboration is established, with the project managed properly from all sides. Progress and impact were accessed regularly. Despite different insights and drivers that each individual organization brought to the table, pivotal decisions were made in a joint manner.
- Create efficient communication and collaboration to allow participants to become comfortable with one another. Doing so accelerated the process of forming, norming, and storming so that the participants were making connections with industry peers and reproducing a work that could be of value to the community.
- Plan effective knowledge transfer by introducing topic related seminars where participants learn recent advances in the area of NLP and experiential learning where participants perform hands-on tasks together, fostering working groups' performance and enhanced productivity.
- Set out experiential learning principles to increase knowledge of industry participants by not only engaging in the experience activity, but also by requiring them to reflect upon their learning and how their skills learned through their project work can be applied beyond the project.
- Vector provided computing cluster access and usage to the industry participants who built their capability on using high performance computers and training large models who would not be able to do otherwise.
- Drive real business impact by providing cutting edge research exposure related to participants' business objectives. Insights gained in the project have informed programs and product development in participating sponsors and provided unique technology development opportunities to industry technical staff.

Conclusion and Future Directions

Through the NLP project, launched in the Summer of 2019, the Vector Institute brought together academic and industry sponsors to explore recent advances in NLP. The primary objectives of this project were to foster and widen collaboration between academic researchers and industry applied scientists on several projects, and to build capacity for further advances and new lines of work in large scale language models in our ecosystem.

The project focused on three areas: domain-specific training, pre-training large models, and the tasks of summarization, question answering, and machine translation.

- **Domain-specific training** where across tasks and explorations confirmed that unsupervised pre-training or BERT in general could improve the performance on fine-tuning tasks. In some tasks, pre-training the general BERT on a legal corpus improved the results over the performance of the base version and the effectiveness of performing domain-specific pre-training on other tasks using finance data created during the project.
- **Pre-training large models** where working groups conclude that $\text{BERT}_{\text{LARGE}}$ pre-training on a cluster of widely available GPUs can be done through careful optimizations and massively parallel workload organization. Working groups also conclude that, given computation resources, the GPT-2 implementation in Tensor2Tensor[34] may still be challenging on a large scale dataset.
- **Summarization, question answering, and machine translation** where working groups demonstrate a clear positive trend in the quality of text summarization for large language models when fine-tuned using domain-specific data (e.g., finance and health data) even in cases where

[34] https://ai.googleblog.com/2017/06/accelerating-deep-learning-research.html

the amount of labeled examples available is relatively small. Although it requires further investigation, using small-sized datasets can be especially useful when the models need to retrained due to (a) data-shift, (b) a wide variety of data-domains, (c) confidential data not publicly available to train the model on, (d) small-size of the domain-specific data available, and (e) a lack of computing resources. In machine translation, the working group observed a significant difference with respect to the best published results of the reference work due to various reasons, including parameter setting and experimental setup. In order to help medical researchers and practitioners to combat pandemic, SIG-Kaggle-COVID19 used its expertise in AI-enabled search capabilities and developed a superior COVID-19 information retrieval system.

Although the project encountered technical challenges such as the appropriateness of datasets, sensitivity and hyper parameters of large-scale models, and project challenges such as lack of time commitment of participants, project meeting expectations, and self-organizing group development and work distribution, with Vector's exclusive collaboration opportunity all participants successfully built their capability on using high performance computers and training large models who would not be able to do otherwise.

The availability of data and computational resources, model sensitivity, and hyperparameterization, as well as the 'black box' nature of these models reveals several potential areas for future work. Among these, naturally, is the continued democratization of these approaches, to ensure that the benefits of modern NLP are made available to all. Increasing the collaborative partnerships, in the way exemplified by this project, will continue to drive real business impact and ensure that opportunities are made available widely.

Appendix

Presentations

- A Partial Replication of Language Representation in the Biomedical Domain, Evolution of Deep Learning Symposium, Poster Presentation, October 2019.
- Multi-Node Training of Large-Scale Language Models, Poster Presentation, Evolution of Deep Learning Symposium, 2019 and SOSP Conference on Operating Systems Principles, AI Systems Workshop, October 2019.
- Multi-Node Training of Large-Scale Language Models, Oral Presentation, GPU Accelerated Conference (GTC) – DC 2019, Washington D.C, November 2019.
- More Questions Than Answers: A Rapid Response to COVID19 Question-Answering, Oral Presentation, Toronto Machine Learning Micro-Summit, April 2020.
- Non-Pharmaceutical Intervention Discovery with Topic Modeling, Poster presentation, ML for Global Health Workshop-ICML'20, July 2020.
- An Experimental Evaluation of Large NLP Models in the Biomedical Domain, Oral Presentation, Vector NLP Symposium, September 2020.
- An Investigation of Transformer-based Language Models in Legal Text, Oral Presentation, Vector NLP Symposium, September 2020.
- FinanceBERT: Sentiment Classification and Extractive Summarization of Financial Text Using BERT, Oral Presentation, Vector NLP Symposium, 2020.
- Multi-node BERT Pretraining: Cost Efficient Approach, Oral Presentation, Vector NLP Symposium, September 2020.
- A Smart System to Generate and Validate Question Answer Pairs for COVID-19 Literature, Oral Presentation, EMNLP 2020 Conference, 1st Workshop on Scholarly Document Processing and Shared Tasks (SDP), November 2020.
- Harnessing the Power of Natural Language Processing (NLP): A Vector Institute Industry Collaborative Project, Oral Presentation, Toronto Machine learning Summit, November 2020.
- Customizing Contextualized Language Models for Legal Document Reviews, the Fourth Annual Workshop on Applications of Artificial Intelligence in the Legal Industry, IEEE Big Data Conference, December 2020.

Participating Contributors List *(in alphabetical order)*

Focus Area 1

Health Domain:

Mohammad Abdalla, Graduate Researcher, University of Toronto, Vector Institute
Saif Charaniya, Sr. Data Scientist, Scotiabank
Paul Grochy, Machine Intelligence Engineer, ROSS
Shobhit Jain, Lead Data Scientist, Manulife
Jin Li, Data Scientist, NVIDIA
Michael Liu, Machine Learning Engineer, Tealbook
Jonathan Lomond, Data Scientist, Intact
Amy Lu, Graduate Researcher, University of Toronto, Vector Institute
Faiza Khan Khattak, Data Scientist, Manulife,
Amandeep Mander, Sr. Data Scientist, Intact
Evan Tam, Data Scientist, Sun Life

Max Tian, Data Scientist, Goldspot Discoveries
Alejandro Troccoli, Sr. Research Scientist, NVIDIA
Kuhan Wang, Sr. Research Scientist, CIBC
Haoran Zhang, Graduate Researcher, University of Toronto, Vector Institute

Finance Domain:

Mahtab Ahmad, Graduate Researcher, Western University, Vector Institute
Stella Wu, Applied Machine Learning Researcher, BMO
Bo Zhao, Applied AI Researcher, BMO

Legal Domain:

Luna Feng, Research Scientist, Thomson Reuters
Borna Jafarpour, Research Scientist, Thomson Reuters
Shohreh Shaghaghian, Sr. Research Scientist, Thomson Reuters

Focus Area 2

Joey Cheng, Machine Learning Research Scientist, Layer 6
Gary Huang, Machine Learning Research Scientist, Layer 6
Thor Johnsen, Deep Learning Software Engineer, NVIDIA
Edward (Zhiyu) Liang, Undergraduate Researcher, University of Toronto, Vector Institute
Jiahuang Lin, Graduate Researcher, University of Toronto, Vector Institute
Purnendu Mukherjee, Deep Learning Software Engineer, NVIDIA
Felipe Perez, Sr. Machine Learning Research Scientist, Layer 6
Filippo Pompilli, Sr. Research Scientist, Thomson Reuters

Focus Area 3

Aisha Alaagib, Research Intern, Vector Institute
Stephane Aroca-Ouellette, Graduate Researcher, University of Toronto, Vector Institute
Nidhi Arora, Sr. Data Scientist. Intact
Akshay Budhkar, Applied Research Scientist, Georgian
Colton Chapin, Machine Intelligence Engineer, ROSS
Arvid Frydenlund, Graduate Researcher, University of Toronto, Vector Institute
Rylan Halteman, Sr. Data Engineer, Wattpad
Matt Kalebic, Data Science Manager, PwC
Faiza Khan Khattak, Data Scientist, Manulife
Brydon Parker, Sr. Data Scientist, Deloitte
Shahin Vakilinia, Sr. Consultant, PwC

SIG-Kaggle-COVID19:

Rohan Bhambhoria, Graduate Researcher, Queen's University
John Chen, Graduate Researcher, University of Toronto, Vector Institute
Conner Cowling, Sr. Research Engineer, Thomson Reuters
Luna Feng, Research Scientist, Thomson Reuters
Borna Ghotbi, Graduate Researcher, University of British Columbia
Hillary Ngai, Graduate Researcher, University of Toronto, Vector Institute
Mah Parsa, Post-doctoral Researcher, University of Toronto, Vector Institute
Dawn Sepehr, Research Scientist, Thomson Reuters
Yoona Park, Graduate Researcher, University of Toronto, Vector Institute,
Jonathan Smith, Machine Learning Scientist, Layer 6
Seung Eun Yi, Machine Learning Scientist, Layer 6

References

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1908.10063.

Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The genetic association database. *Nature genetics*, *36*(5), 431-432.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bhambhoria, R., Feng, L., Sepehr, D., Chen, J., Cowling, C., Kocak, S., & Dolatabadi, E. (2020, November). A Smart System to Generate and Validate Question Answer Pairs for COVID-19 Literature. In *Proceedings of the First Workshop on Scholarly Document Processing* (pp. 20-30).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1810.04805.

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of biomedical informatics, 47, 1-10.

Drysdale, E., Dolatabadi, E., Chivers, C., Liu, V., Saria, S., Sendak, M., Wiens, J., Brudno, M., Hoyt, A., Mazwi, M., & Others. (2019). Implementing AI in healthcare. https://vectorinstitute.ai/wp-content/uploads/2020/03/implementing-ai-in-healthcare.pdf.

Du, Y., Li, Q., Wang, L., & He, Y. (2020). Biomedical-domain pre-trained language model for extractive summarization. Knowledge-Based Systems, 105964.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, *28*, 1693-1701.

Huang, X. S., Perez, F., Ba, J., & Volkovs, M. (2020, November). Improving transformer optimization through better initialization. In *International Conference on Machine Learning* (pp. 4475-4483). PMLR.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). PubMedQA: a dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146.

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. Journal of Biomedical Informatics: X, 4, 100057.

Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., & Collier, N. (2004, August). Introduction to the bio-entity recognition task at JNLPBA. In Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (pp. 70-75).

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1909.11942.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics , 36(4), 1234–1240.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lin, J., Li, X., & Pekhimenko, G. (2020). Multi-node Bert-pretraining: Cost-efficient Approach. *arXiv preprint arXiv:2008.00177*.

Lin, H., & Ng, V. (2019, July). Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 9815-9822).

Liu, Y. (2019). Fine-tune BERT for Extractive Summarization. In arXiv [cs.CL]. arXiv.

http://arxiv.org/abs/1903.10318.
Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. arXiv preprint arXiv:1908.08345.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1907.11692.
Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782-796.
Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., & Balahur, A. (2018, April). WWW'18 open challenge: financial opinion mining and question answering. In Companion Proceedings of the The Web Conference 2018 (pp. 1941-1942).
Pafilis, E., Frankild, S. P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., ... & Jensen, L. J. (2013). The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. PloS one, 8(6), e65390.
Phan, M. C., Sun, A., & Tay, Y. (2019, July). Robust representation learning of biomedical names. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3275-3285).
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.
Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.
Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.
Shaghaghian,S., Feng, L., Jafarpour , B., & Pogrebnyakov, N. (20020). Customizing Contextualized Language Models for Legal Document Reviews. The Fourth Annual Workshop on Applications of Artificial Intelligence in the Legal Industry, IEEE Big Data Conference. 2020.
Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.
Smith, J., Ghotbi, B., Yi, S., & Parsapoor, M. (2020). Non-Pharmaceutical Intervention Discovery with Topic Modeling. arXiv preprint arXiv:2009.13602.
Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. arXiv preprint arXiv:1905.02450.
Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., ... & Almirantis, Y. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics, 16(1), 138.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. U., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 (pp. 5998–6008). Curran Associates, Inc.
Völske, M., Potthast, M., Syed, S., & Stein, B. (2017, September). Tl; dr: Mining reddit to learn automatic summarization. In Proceedings of the Workshop on New Frontiers in Summarization (pp. 59-63).
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 32 (pp. 5753–5763). Curran Associates, Inc.

Yogatama, D., de Masson d'Autume, C., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., & Blunsom, P. (2019). Learning and Evaluating General Linguistic Intelligence. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/1901.11373.