Optimizing Inference Performance of Transformers on CPUs

Dave Dice
Oracle Labs
Burlington, MA, USA
dave.dice@oracle.com

Alex Kogan
Oracle Labs
Burlington, MA, USA
alex.kogan@oracle.com

ABSTRACT
The Transformer architecture revolutionized the field of natural language processing (NLP). Transformers-based models (e.g., BERT) power many important Web services, such as search, translation, question-answering, etc. While enormous research attention is paid to the training of those models, relatively little efforts are made to improve their inference performance. This paper comes to address this gap by presenting an empirical analysis of scalability and performance of inferencing a Transformer-based model on CPUs. Focusing on the highly popular BERT model, we identify key components of the Transformer architecture where the bulk of the computation happens, and propose an Adaptive Linear Module Optimization (ALMO) to speed them up. The optimization is evaluated using the inference benchmark from HuggingFace, and is shown to achieve the speedup of up to x1.71. Notably, ALMO does not require any changes to the implementation of the models nor affects their accuracy.

1 INTRODUCTION
The introduction of the Transformer architecture for deep neural networks (DNNs) by Vaswani et al. [27] has literally transformed the field of NLP. It happened just a few years ago (in 2017, to be exact), and since then the field has exploded with an enormous wave of Transformer-based models achieving state-of-the-art, and often super-human, performance on many NLP tasks, which just recently have been considered unrealistically difficult to solve. BERT [5], RoBERTa [13], ALBERT [11], Transformer-XL [4] are only very few examples in the vast sea of published models [36]. As of today, Transformer-based models, and BERT in particular, power many important Web services, such as search [15, 16], translation and text classification [12].

The big premise of the Transformer-based models is that they can be pre-trained on huge amounts of unlabeled data (such as all of Wikipedia or a book corpus), and later fine-tuned to a specific task (e.g., question-answering) using just a small amount of labeled data. To achieve high accuracy, those models feature millions (and, at times, billions) of parameters, and require long and expensive training. As a result, numerous efforts have been made to optimize the training performance of those models [8, 11, 13, 37]. At the same time, and despite the vast deployment of those models in practice, far less attention is paid to inference performance. Furthermore, among the efforts that do target inference performance of Transformer-based models, many consider GPU or smartphone-based deployments [7, 30, 34, 39], even though in many practical settings the inference is done on small CPU-based systems [12, 35].

This paper comes to address this gap by presenting an empirical analysis of scalability and performance of inferencing Transformer-based models on CPUs. We identify the key component of the Transformer architecture where the bulk of the computation happens, namely, the matrix multiplication (matmul) operations, and propose three optimizations to speed them up. Due to space constraints, only one of those optimizations is described in this short paper; the other two optimizations are discussed in the longer version [6].

The first optimization, which we call ALMO and describe in detail in this paper, is based on the observation that the performance of the matmul operation is heavily impacted not only by the shape (dimensions) of the source matrices and the available computing resources (the number of worker threads), but also by whether (at least) one of those matrices is provided in a transposed form. We propose a lightweight method to adaptively choose the appropriate form of source matrices for the inference, which results in substantial performance improvement of the latter. The second optimization stems from the observation that an invocation of matmul operations in deep learning (DL) frameworks incurs a significant sequential overhead, leading to the poor scalability of those operations. We analyze the source of the overhead, and demonstrate how the scalability of matmul operations (and the overall inference performance) can be improved by reducing (some portion of) that overhead. Finally, the third optimization builds on the realization that while performant matmul operations are typically implemented by partitioning matrices into sub-matrices and carrying out the actual computation using highly optimized inner kernels [9], the partitioning itself might be suboptimal and not fully utilize parameters of the underlying hardware (such as cache capacity). We show how choosing different parameters for matrix partitioning results in faster matmul operations. We evaluate the efficacy of our optimizations using the industry-grade inference benchmark from HuggingFace [33].

We note that prior work shows many factors impacting the inference performance of DNN models [32], including the choice of a DL framework, a math library, a thread pool library, availability of certain hardware features, such as the support for SIMD (single instruction multiple data), etc. To make our analysis feasible, we make several conscious choices when setting up our experimentation environment, focusing on the inference performance of BERT implemented in the widely used Pytorch framework [21] built with the state-of-the-art oneDNN math library [19] (previously known as MKL-DNN) and run on an Intel Skylake processor-powered system (which supports AVX512 SIMD instructions). While the chosen setup is significant, we validate the generality of many of our findings in other variations of our setup such as with other Transformer-based models and math libraries. We also note that despite the focus of our work being on NLP and Transformer-based models in particular, we believe our findings extend to any model...
in which matmul operations consume a significant portion of the inference time.

The rest of the paper is organized as follows. We provide the relevant background on Transformers and BERT in Section 2. The related work is discussed in Section 3. We describe our evaluation setup in Section 4 and provide the analysis of the inference performance of BERT on CPUs in Section 5. Based on this analysis, we describe the Adaptive Linear Module Optimization (ALMO) for inference performance in Section 6; additional optimizations are described in the longer version of this paper [6]. Finally, we conclude in Section 7 with a discussion of the results and some of the future directions.

2 BACKGROUND: TRANSFORMER AND BERT

The Transformer architecture [27] is composed of two stacks of identical layers; those stacks are called encoder and decoder. For the purpose of this paper, we will focus on the encoder stack only, which is used exclusively in many actual Transformer-based models, including BERT. In fact, we will note upfront that BERT’s model architecture is almost identical to the Transformer encoder, only tweaking the number of layers, the activation function, etc. [5]. Also, we note that BERT itself has multiple configurations that differ in the various model hyper-parameters (e.g., the “base” configuration for BERT has 12 layers while the “large” one has 24). Unless specified otherwise, when we say BERT in this paper, we refer to its “base” configuration [5].

Each encoder layer has two sublayers, the first being a multi-head self-attention mechanism and the second being a position-wise fully-connected feed-forward network. A residual connection is employed around each of the sub-layers, followed by a layer normalization.

The attention mechanism is at the heart of the Transformer architecture. For the purpose of this paper we will focus on the actual computations performed by this mechanism; the explanation of the intuition behind those computations can be found in many excellent sources [1, 17], including in the original paper [27]. Specifically, the attention mechanism takes as an input three matrices Q, K and V and computes the output matrix:

\[
\text{Attn}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V
\]

where \(d_k\) is the attention input dimension (64 for the BERT model). As mentioned above, each self-attention sublayer includes multiple heads (12 for the BERT model). The computed function of this sublayer is given by the following expressions:

\[
\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)WO
\]

where \(\text{head}_i = \text{Attn}(QW_i^Q, KW_i^K, VW_i^V)\)

where \(WO, W_i^Q, W_i^K, W_i^V\) are parameter matrices. Overall, the computation of the multi-head self-attention requires 4 matrix multiplications to create input token projections (the Q, K and V matrices) and the projection of the concatenated output of all the multiple heads. (We note that when Transformer is implemented in Pytorch, each of those multiplications are performed during the computation of the corresponding Linear modules.) In addition, two batched matrix multiplications are required to calculate the \(\text{Attn}\) function in Equation 1. Furthermore, the self-attention sublayer includes the invocation of softmax and layer normalization operations.

As for the fully-connected feed-forward sublayer, it consists of two linear transformations with an activation function in between:

\[
\text{FFN}(x) = \text{Act}(xW_1 + b_1)W_2 + b_2
\]

where \(W_1, b_1, W_2\) and \(b_2\) are weight and bias matrices, respectively (which are model parameters, one set for each layer) and \(\text{Act}\) is an activation function, such as gelu [10]. While the inputs and outputs of the feed-forward sublayer have the same dimensions as the rest of the model (768, in case of BERT), the inner-layer has a larger dimensionality (3072 for BERT). It is easy to see that the computation of the feed-forward sublayer requires two matrix multiplication operations (carried by two Linear modules in Pytorch), as well as an activation function and a layer normalization operation.

3 RELATED WORK

There is a relatively small body of work we are aware of on optimizing inference performance of NLP models on CPUs. Ning et al. [16] describe their effort on accelerating BERT with ONNX Runtime, an inference engine compatible with PyTorch and TensorFlow. The idea is to fuse multiple operations in the computation graph (e.g., matrix multiplication, layer normalization and gelu) to reduce the amount of overhead (e.g., memory copying) in invoking each elementary computation individually. They also experiment with reducing the number of layers in BERT sacrificing (some) accuracy for higher performance. In general, we note that the operation fusion is a known technique for optimizing inference performance, and is orthogonal to the optimizations described in this paper. Also, our techniques aim for performing the given inference computations faster, but without any change to the accuracy.

Wu et al. [35] describe another effort to optimize inference of BERT in Apache MXNet using the GluonNLP toolkit. They report on speedups achieved by using the MKL math library in MXNet as well as from quantizing the model for better performance with lower precision. We note that we use MKL as one of the baseline configurations in our analysis, while model quantization is, once again, orthogonal to the ideas discussed in this paper and may result in reduced accuracy.

There is an enormous effort on refining and/or replacing the attention mechanism with a more efficient alternative that requires less computation and/or allows scaling for longer sentences, e.g., [2, 3, 28, 38]. While most of those efforts are primarily concerned with speeding up training, they help inference directly or indirectly as well. Notably, one of the goals behind the knowledge distillation effort [25, 26, 29], i.e., training a smaller model (student) to achieve a similar accuracy as a larger one (teacher), is reducing the inference latency. Indeed, Le and Kaehler describe how they employ distillation with quantization to speedup their deployment of BERT on CPUs [12]. We believe the optimizations described in this paper apply to most of such work. In particular, we show the speedups achieved by our optimization for inferring Distil-Bert [25], a popular model that uses knowledge distillation, are similar to those of BERT.
Optimizing Inference Performance of Transformers on CPUs

(a) Seq. length 8
(b) Seq. length 64
(c) Seq. length 384

Figure 1: Inference latency and the breakdown of runtime spent in different sub-layers (Self-Attention, Feed-Forward, and their components) for BERT and input sequences of various sizes.

In a broader context, Liu et al. [14] describe an approach called NeoCPU for optimizing CNN inference on CPUs. In addition to the common optimizations of operation fusion and inference simplification, NeoCPU manipulates the data layout flowing through the model to minimize the overhead of transforming the data between various individual operations. Fang et al. [7] present TurboTransformers, a GPU-based serving system for Transformer models. They describe a number of optimizations targeting GPUs, such as memory management and batch reduction. Optimizing inference of NLP models on GPUs has been also the motivation behind the work by Wang et al. [30]. At the same time, Wu et al. [34] describe opportunities and design challenges in enabling machine learning inference on smartphones and other edge platforms.

4 EVALUATION ENVIRONMENT

In this section, we describe the hardware and software setup for our experiments. We ran the experiments on an Intel-based system featuring two Intel Xeon Platinum 8167M processors with 26 hyper-threaded cores each, and runs an Oracle Linux 7.8 OS. To avoid any non-uniform memory access effects, we use the `numactl` utility to restrict all our experiments to executing on and allocating memory from one socket only.

On the software side, we use Pytorch v1.6, a popular DL framework. We compile Pytorch in the default configuration, which means that it employs MKL as its default math library, but also includes support for oneDNN. While MKL is a closed-source library, oneDNN is “an open-source cross-platform performance library of basic building blocks for deep learning applications” [19], and, unless stated otherwise, we use the latter in our experiments.

To invoke oneDNN bindings, one needs to convert a given model as well as the input tensors into the so-called “mkldnn” data format, which dictates how data is laid out in memory [20]. We note that oneDNN bindings, however, are available for only a handful of DL operations, such as Linear and Convolution modules. In practice, this means that for each such supported operation, Pytorch would seamlessly convert input tensors into the mkldnn format, apply the corresponding operation and convert the output tensors back into the default (“dense”) format so they could be sent to other operations for which oneDNN bindings are not provided. Such conversions do not come for free, however, and involve memory layout translations and memory copying. To avoid this overhead, we extend the integration of oneDNN with Pytorch, adding the missing bindings for various operations invoked by a typical Transformer model, such as the layer normalization, softmax and gelu activation functions. The extension comprises of a few hundred lines of C++ and Python code. As we show in Section 6, the resulting setup performs on-par with or better than the default Pytorch configuration (which employs MKL).

We use the popular Transformers Python package (v3.0.2) from HuggingFace, which provides a state-of-the-art implementation of numerous Transformer-based NLP models implemented in Pytorch (and Tensorflow) [33]. In addition, Transformers includes an easy-to-use inference benchmark, which we utilize heavily for our experiments. Furthermore, we utilize mcbench [31], the opensourced suite of microbenchmarks, which includes, among other things, microbenchmarks for evaluating the performance of matrix multiplication operations when invoked directly through the C++ API of the corresponding math libraries.

5 INFERENCE PERFORMANCE ANALYSIS

We instrument the BERT model implemented in Transformers [33] and collect timing information for various sub-layers (multi-head attention, feed-forward) and modules (Linear, Softmax, etc.) composing the model while executing the Transformers inference benchmark. We experiment with various input sequence lengths and vary the number of threads from 1 to 16 (by setting the `OMP_NUM_THREADS` environment variable). We note that although our experimental machine has more than 16 cores, we deliberately decided to focus on smaller setups, as practical inference deployments typically include a small number of cores [12, 35].

Figure 1 presents the inference latency as well as the breakdown of runtime spent in two main sub-layers, attention and feed-forward (along with the small portion of time not associated with any of the sub-layers, which consists mostly of input embedding and pooling of the resulting tensor; this time is denoted as a red box titled “Other”). In addition, we break the time in the two sub-layers into major components. For the attention sub-layer, this is the time spent...
we note that the overall scalability of the inference latency is relative. While the portion of time spent in matmul operations shrinks. The attention sub-layer does include other operations (including layer normalization, softmax, tensor transpose and reshaping operations, etc.), which explains why the time share of the attention sub-layer grows as we increase the number of threads. Specifically, as we show in the longer version of our paper [6], matmul operations, being carried out by carefully optimized math libraries (oneDNN, in this case), scale almost linearly with the number of threads. At the same time, other operations (including layer norm, softmax and tensor reshaping) do not scale. As such, their relative portion grows while the portion of time spent in matmul operations shrinks. The portion of non-scalable operation is larger in the attention sub-layer, hence its weight grows with the number of threads. This growth is more pronounced when the input sequence is large (cf. Figure 1 (c)) since the matmul operations are invoked with larger matrices, thus consuming a larger portion of time w.r.t. all other operations.

When examining the total runtime (the black curve in Figure 1), we note that the overall scalability of the inference latency is relatively low, and depends on the input sequence length. In particular, for sequences of 8 tokens, we achieve the speedup of only x3.3 when running with 16 threads versus 1 thread; the speedup goes up to x9.7 for sequences of 384 tokens. Better scalability with longer sequences is, once again, related to the scalability of matmul operations in math libraries and the fact that larger sequences result in heavier matmul operations (with larger matrices) — reducing the time spent in matmul operations when the number of threads increases has a larger effect on the overall scalability of the inference latency.

In Figure 2, we present a different way to breakdown the inference runtime, by the time spent in various models. While most module names are self-explanatory, we note that “bmm” stands for batched matrix multiplication, the operation at the heart of the multi-head attention mechanism (there are two of those operations per each attention layer in BERT); “other” stands for the time spent in computations not included in the specific modules, such as the time spent on transposing and reshaping tensors, input embedding, pooling of the resulting tensor, etc.

Not surprisingly, the vast majority of the inference time is spent in the Linear module, which in practice means matmul operations (the Linear module applies a linear transformation to the incoming data, calculating a product of the input matrix with the stored weight matrix). This concurs the results in Figure 1. When summing up both “linear” and “bmm” runtime portions, the matmul operations consume between 66.2% and 91.5% of the total runtime. Note that this portion decreases as we increase the number of threads. As explained above, this is because matmul operations are executed by a math library (oneDNN), and are highly optimized and scale almost linearly with the number of threads.

6 ADAPTIVE LINEAR MODULE

The results in Section 5 show that the key to improving inference performance on CPU lies in reducing the time spent in matmul operations.

For the context, the API for a matmul operation allows invoking that operation on two source matrices A and B (producing the destination matrix C=AB) s.t. each of the source matrices can be provided in the transposed form [18]. At the same time, the Pytorch Linear module stores the weight matrix in a transposed form, which means that, during inference, the input matrix (A) is always non-transposed, while the weight matrix (B) is always transposed. We believe the Pytorch designers made this choice (to transpose the weight matrix) to achieve better performance for the backward pass during training [22], a concern which is not relevant for inference.

Our experiments with the matmul microbenchmark from mcbench reveal an interesting observation. Figure 3 demonstrates the ratio between the time to compute matmul when both source matrices are non-transposed to the time to compute matmul when (only) the second source matrix is transposed. In other words, ratio > 1 (ratio < 1) corresponds to cases in which the former (latter, respectively) method is faster. The shape of the second matrix (B) is represented by the name of the corresponding data series, while the shape of the first matrix (A) is given by the sequence length x first dimension of B. Note that the chosen three shapes are not

---

1We note that in Tensorflow, the weight matrix is always given in the normal form to the matmul operation.
Optimizing Inference Performance of Transformers on CPUs

Figure 3: Matmul operation performance as a ratio between the time to multiply two non-transposed matrices and the time to multiply a non-transposed matrix by a transposed one. That is, ratio > 1 corresponds to cases in which the former method is faster, and vice versa.

Incidental, and they correspond to the shapes of weight matrices used in Linear modules of the BERT model.

More concretely, Figure 3 (a)–(d) compare the performance of the matmul operation in oneDNN across different numbers of threads. We see that for shorter sequences, multiplying the non-transposed matrices is almost always faster, and often results in substantial speedups. For longer sequences, the picture is less clear – one way of applying a matmul operation is faster than the other for one shape but worse for another. In general, the faster way of applying a matmul operation depends on the shape of the source matrices and the number of threads used. We also confirmed that this observation is not unique to oneDNN, and is reproducible, at least to some extent, with other math libraries. Figure 3 (e) and (f) show the results obtained with MKL and OpenBLAS libraries, respectively (for the latter, we used the benchmark included in the library sources; for brevity, we include only the result for 16 threads).

One may wonder about the reason for this performance difference. In oneDNN, much like in any math library for high-performance matmul calculation [9], the matmul operation is coded in assembly, and each of the matmul variants (e.g., one with both source matrices in the normal form vs. one in which the second matrix is transposed) results in a different code path, which generates different memory access patterns. Based on the profiling information produced by perf, we observe that given a certain configuration (i.e., the same source matrix shapes and the number of threads), both variants have a similar number of L1 data cache misses, but the faster variant has a lower number of L3 cache accesses. This suggests that one reason for performance difference might be the better utilization of L2 cache by one variant over the other.

Given the results in Figure 3, we propose the following optimization for the Linear module. Each Linear module is augmented with a transposeFlags array, specifying whether to use a transposed version of the weights matrix for the forward pass (inference). Entry $i$ of the array corresponds to the sequence length of $2^i$; the array has 10 entries corresponding to the maximal length of 512 tokens. When creating a Linear module with the given weights shape $[in, out]$, we generate random matrices with the shape $[2^i, in]$, for each $0 \leq i < 10$, and measure the time to perform a matmul operation when the weight matrix is transposed or not. Based on the result, we set the corresponding entry transposeFlags[$i$]. During the inference time, given the input of shape $[length, in]$, we calculate $s = \lfloor \log(length) \rfloor$, and based on the flag in transposeFlags[$s$], perform the matmul operation with either weight matrix transposed or not.

To avoid the overhead of transposing the weight matrix during inference, we keep both variants of the weight matrix (transposed and non-transposed one). This doubles the memory footprint of the Linear module. While it might not be a concern on some CPU deployments, there are several ways to mitigate this drawback. First, some shapes always prefer one form over the other, for all thread counts (e.g., the shape 3072-768 in Figure 3 (a)–(d)). For this case, we can keep only the relevant variant of the weight matrix. Second, the length of the input can be known prior to the deployment of an inference server, e.g., in a farm of inference servers, certain servers can be configured to handle input of a certain sequence length. Once again, in this case we can keep only the relevant variant of the weight matrix. Finally, if the input range is dynamic, one can store one variant of the weight matrix and transpose on-demand. The selection of the stored variant can be also dynamic and tuned based on the actual input lengths seen during the runtime. All those mitigation ideas are left for the future work.

We note that transposeFlags can be shared among Linear modules of the same shape. We use a key-value map (dictionary) to store
Table 1: BERT-base inference latency (ms), for various sequence lengths. The numbers in () show the speedup of onednn-almo over onednn-base. The numbers after the ± sign specify the standard deviation when it is larger than 1% of the mean.

Table 2: RoBERTa inference latency (ms).

Table 3: DistilBERT inference latency (ms).

According to Figure 3, multiplying the transposed weight matrix is faster. The adaptive variant selects the correct shape in each case, performing on-par or better than the other two onedNN-based variants.
One future direction for this work is scaling primitive operations beyond matmul. While matmul is responsible for the lion’s share of the inference time, the portion of other operations grows as the number of threads increases. For instance, for short sequences, the share of the time spent in the layer normalization operation grows from 2.9% for 1 thread to 12% for 16 threads (cf. Figure 2). Parallelizing those operations (and potentially fusing them with matmul) should provide further improvement to the inference performance.

REFERENCES