

Improving the Efficiency of Transformers for Resource-Constrained Devices

Hamid Tabani*, Ajay Balasubramaniam, Shabbir Marzban, Elahe Arani, Bahram Zonooz
Advanced Research Lab, NavInfo Europe, Eindhoven, The Netherlands
{firstname.lastname}@navinfo.eu, bahram.zonooz@gmail.com

Abstract—Transformers provide promising accuracy and have become popular and used in various domains such as natural language processing and computer vision. However, due to their massive number of model parameters, memory and computation requirements, they are not suitable for resource-constrained low-power devices. Even with high-performance and specialized devices, the memory bandwidth can become a performance-limiting bottleneck. In this paper, we present a performance analysis of state-of-the-art vision transformers on several devices. We propose to reduce the overall memory footprint and memory transfers by clustering the model parameters. We show that by using only 64 clusters to represent model parameters, it is possible to reduce the data transfer from the main memory by more than 4x, achieve up to 22% speedup and 39% energy savings on mobile devices with less than 0.1% accuracy loss.

Index Terms—Deep Learning, Transformers, Clustering, Resource-Constrained Devices

I. INTRODUCTION

Transformers, deep neural network architectures based on self-attention mechanism, were developed to solve the problem of sequence understanding [1]. Transformers are widely used for different applications. For instance, they have been used by OpenAI in their language models and by DeepMind in AlphaStar. Although they garnered popularity in other domains such as computer vision [2], they come at the cost of having huge number of model parameters.

Therefore, training and inference of transformers are computation-intensive resulting in massive memory usage and energy consumption. BERT (Bidirectional Encoder Representations from Transformers) [3], is a transformer-based machine learning model for NLP developed by Google. It has been shown that speeding up the training of BERT by 30% translates into a savings of over \$85,000 on Amazon Web Service (AWS) [4]. For the GPT-3 transformer model [5] with a training cost of \$12M, this 30% speedup could save \$3.6M and more than 120 MWh energy [4]. The training process is usually done once using powerful machines while the inference process of the trained model can be deployed many times on resource-constrained devices such as wide range of mobile and wearable devices. Hence, in this paper, we focus on optimizing transformers for inference. Note that it is very challenging to significantly reduce the energy consumption and memory usage of transformers while maintaining the accuracy intact. On the other hand, efficient solutions may

only be feasible with adequate hardware support. For instance, to perform the arithmetic operations in 16-bit floating-point (FP16) instead of 32-bit floating-point (FP32), we require the hardware to support FP16 arithmetic.

Benefits of transformers can be leveraged by deploying them on resource-constrained devices. However, on such devices, memory transactions can become the main bulk of energy consumption. In recent years, application-specific accelerators have become more and more popular due to their efficient performance which make them suitable for low-power mobile and wearable devices. In fact, most of today's devices are integrating tens of accelerators for different tasks [6]. Specialized accelerators and hardware platforms are designed to benefit from application-specific features such as parallelism, quantization, and sparsity due to pruning [7]. However, in some cases, the peak performance cannot be achieved due to the memory bandwidth saturation. This is due to the fact that tens of Gigabytes of data needs to be fetched from the main memory and processed in a fraction of a second. This can get worse in heterogeneous system-on-chips (SoC) where multiple computing units such as multicore CPUs, GPUs, and other accelerators access the same memory subsystem simultaneously [8], [9].

With this understanding about resource-constrained devices, we believe that the efficiency of transformers can be significantly improved. Thus, we aim at further improving the performance and energy consumption of transformers for such devices. Our main contributions are as follows:

- 1) We analyze transformers for inference, identifying key functions which are the most time- and energy-consuming.
- 2) Applying clustering schemes on state-of-the-art transformer models for computer vision which reduces the overall size of the parameters and directly impacts the memory bandwidth.
- 3) We show that significant energy and performance improvements can be achieved with negligible impact on the accuracy of the models. We present the existing trade-offs and opportunities to maximize the utilization of specialized accelerators.

The rest of the paper is organized as follows: Section II presents background on transformers. Section III introduces our analysis and approach towards optimization of transformers and the proposed techniques. In Section IV, we present our

* Corresponding Author. This paper is accepted as a full paper at 24th Euromicro Conference on Digital System Design (DSD).

methodology and experimental setup. In Section V, we present the results and discuss the existing trade-offs. Section VI summarizes the related work. Finally, Section VII concludes the paper.

II. BACKGROUND

In this section, we present a brief background on Transformers, their architectural design and features, and their advantages over previous models.

Before the introduction of Transformers, most state-of-the-art NLP systems relied on gated recurrent neural networks (RNNs) [11], such as LSTMs [12] and gated recurrent units (GRUs) [13], with added attention mechanisms. The Transformer architecture is based on these attention technologies but does not use an RNN framework, demonstrating that attention mechanisms alone, without recurrent sequential processing, are powerful enough to achieve RNN-like efficiency.

Tokens are processed sequentially by gated RNNs, which keep a state vector that contains a representation of the data seen after each token. To process the n th token, the model creates a new state that represents the sentence up to token $n - 1$ by combining the state representing the sentence up to token $n - 1$ with the information from the new token [13]. The knowledge from a single token will potentially spread forever.

The implementation of attention mechanisms helped to solve this issue. Attention mechanisms allow a model to look at and draw from the state of the sentence at any point in time. The attention layer has access to all previous states and weighs them according to a learned measure of relevance to the current token, allowing it to provide more precise details about distant related tokens. Translation is an excellent example of the value of focus. The first word of the French output is almost certainly influenced by the beginning of the English input in an English-to-French translation scheme. In a typical encoder-decoder LSTM model, however, the model is only given the state vector of the last English word to generate the first word of the French output. Theoretically, this vector can store information about the entire English sentence, providing the model with all required information; however, in practice, this information is frequently lost. When an attention mechanism is added to the model, it will learn to pay attention to the states of early English tokens when generating the beginning of the French production, giving it a much better understanding of what it is translating.

The same issue that affects RNNs in general also affects LSTMs, namely that when sentences are too long, LSTMs perform poorly. The explanation for this is that the likelihood of remembering the meaning of a word far away from the actual word being processed decreases exponentially with time. As a consequence, when sentences are long, the model often forgets the content of positions further down the sequence. Another issue with RNNs and LSTMs is that they are difficult to parallelize for processing sentences since they must be processed word by word. Furthermore, there is no paradigm for long and short-term dependencies. To summarize, LSTMs and RNNs have three issues: (1) Parallelization is hampered by

sequential computation, (2) Long and short-term dependencies are not explicitly modeled, and (3), The “distance” between positions is a straight line.

When attention mechanisms were applied to RNNs, they resulted in significant performance improvements. The Transformer’s introduction revealed that attention mechanisms were effective in and of themselves, and that sequential recurrent data processing was not needed to achieve the performance gains of RNNs with attention. The Transformer, instead of being an RNN, uses an attention system that processes all tokens at the same time and calculates attention weights between them. Transformers can be trained more effectively on larger datasets because they do not rely on sequential processing and lend themselves easily to parallelization.

Transformers are multi-layered structures made up of Transformer blocks stacked on top of each other. A multi-head self-attention mechanism, a position-wise feed-forward network, layer normalization modules, and residual connectors distinguish transformer blocks [14]. Convolutional Neural Networks (CNNs) have inductive biases including translation invariance and a locally limited receptive region, which transformers ignore. Invariance refers to the ability to recognise an entity (i.e. object) in a picture through changes in its presence or location. In computer vision, translation means that each image pixel has been shifted in a certain direction by a specified number. Note that convolution is a linear local operator. Only the neighbor values, as shown by the kernel, are available. The transformer, on the other hand, is permutation invariant by nature and therefore cannot process grid-structured data. Therefore, a spatial non-sequential signal is translated to a series for this purpose.

Vision transformer is the name of the overall architecture (ViT in short). It first creates patches from a picture and gets the patches as flat as possible ①. Then, it convert the flattened patches into lower-dimensional linear embeddings ②. Next, positional embeddings should be added to the sequence and be fed as an input to a standard transformer encoder ③. Image labels are used to pre-train the model in a fully supervised manner using a huge dataset. Finally, image detection should be fine-tuned on the downstream dataset. Figure 1 shows how Visual Transformers (ViT) are a straightforward application of the transformer architecture in image classification [10].

III. PERFORMANCE ANALYSIS AND PROPOSED OPTIMIZATIONS

In this section, we first present the performance analysis of transformers and then, we present the clustering techniques we applied on transformers for both GPUs and accelerators.

A. Performance Analysis

In our experiments, we use two of the latest transformer models for classification:

- Classification Transformer (ViT) from Google [10] which achieves excellent results compared to state-of-the-art convolutional networks.

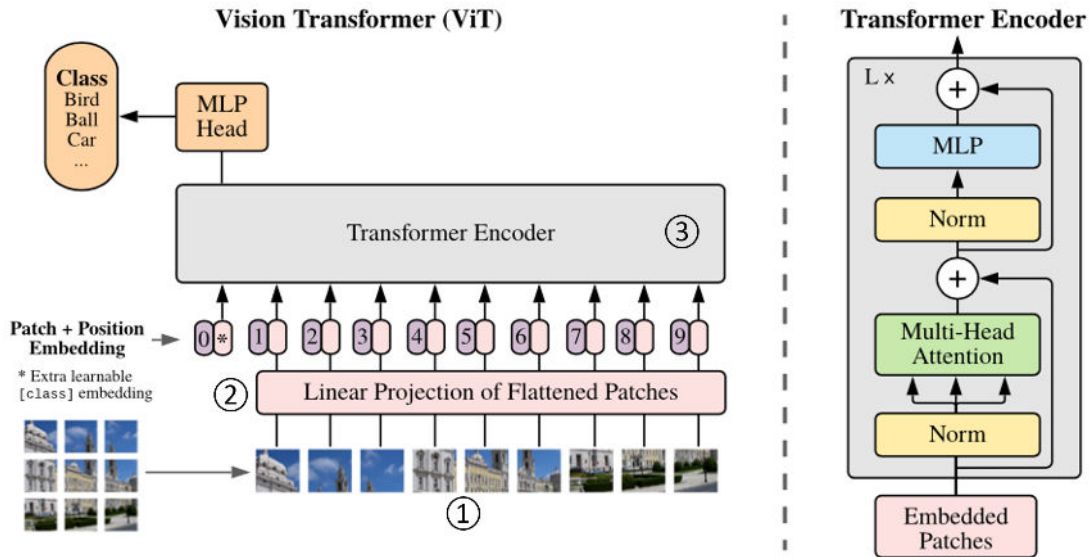


Fig. 1. Application of the transformer architecture to image classification [10].

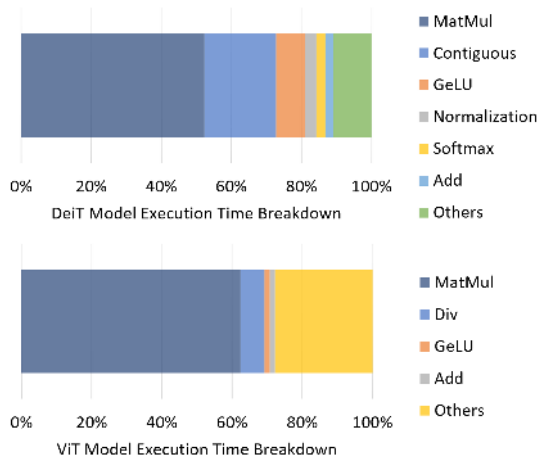


Fig. 2. Execution time breakdown of the DeiT and ViT models.

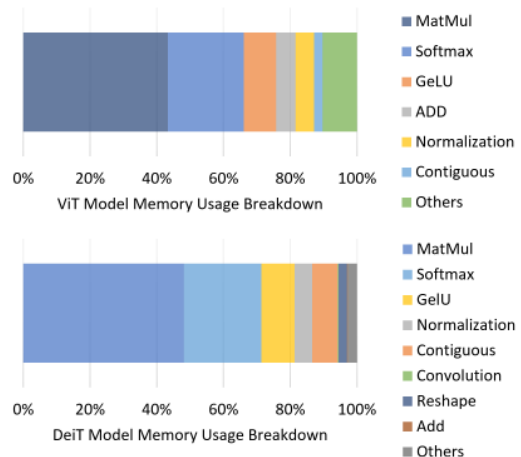


Fig. 3. Memory usage breakdown of the DeiT and ViT models.

- DeiT: Data-efficient Image Transformer from Facebook research, a state-of-the-art vision transformer [15] with 86M parameters.

We first profile the two classification models running on the high-end NVIDIA 2080 Ti GPU. The execution time breakdown of different functions of each model running on the GPU is shown in Figure 2. Matrix multiplication is one of the most time-consuming processes during an inference of both models taking more than 50% of the execution time.

We have analyzed the memory usage for each model as it is shown in Figure 3. As the numbers show, in both models, the parameters to perform the matrix multiplication operations are taking more than 40% of the memory. Next, Softmax and other layers are the most memory-demanding operations as

Figure 3 shows.

B. Clustering

For both efficiency and energy consumption, reducing the size of the parameters of deep learning models is critical. The efficiency of inference in transformers is primarily limited by memory transfers. Furthermore, since off-chip memory accesses are one of the main sources of energy consumption in mobile devices, reducing off-chip memory accesses result in significant energy savings [16].

Clustering model parameters using K-means [17] algorithm is one of the most effective techniques of non-linear quantization for parameter compression. Most popular approaches use nonlinear quantization systems, such as K-means, to minimize data size by up to 4 times with negligible change in accu-

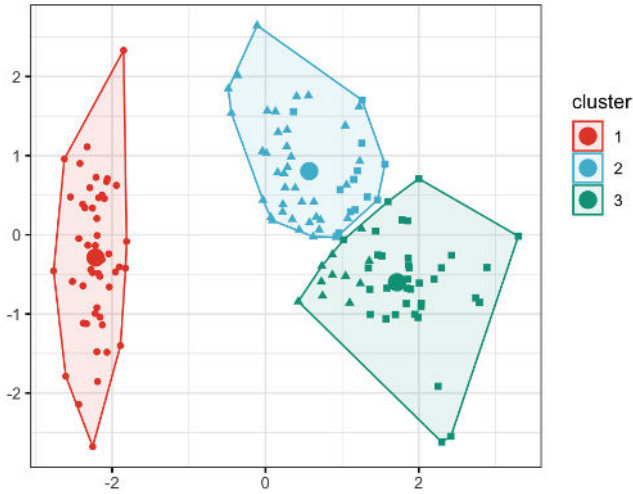


Fig. 4. K-means clustering algorithm grouped the data into three clusters each with a centroid.

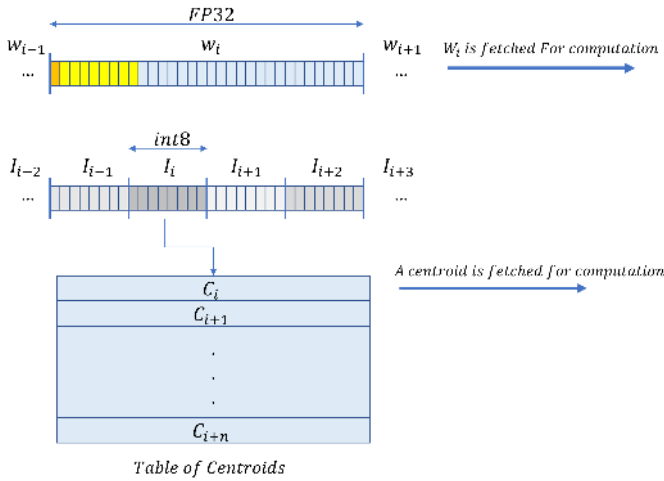


Fig. 5. The process of using floating-point 32-bit default parameters vs. using clustered parameters.

racy [18], [19]. By applying K-means on model parameters, we group the parameters into k clusters each of which has a cluster *centroid*. The centroids are stored in a table referred to as “table of centroids”. Then, in the step of non-linear quantization, we replace all weights of every cluster by the representative indices from the table of centroids as shown in Figure 4.

Clustering replaces floating-point values in a codebook of centroids with slightly smaller integer indices. When using K-means with 256 clusters, for example, each 32-bit floating-point (FP) value is replaced by an 8-bit index, resulting in a compression ratio of nearly 4x (only 1 KB is required for the storage of a table of 256 32-bit centroids). To use the clustered parameters, the corresponding 8-bit index is fetched instead of the 32-bit parameter in the baseline. The index is used to pick the corresponding *centroid* from the very small table of centroids as shown in Figure 5.

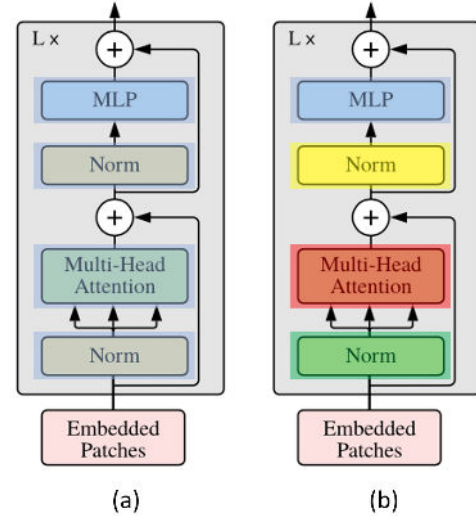


Fig. 6. Clustering (a) entire model parameters vs. (b) per-layer clustering.

In this paper, we explore scalar clustering in which every single parameter in the model will be directly represented by an index. We cluster the model parameters in two ways, as shown in Figure 6:

- 1) **Clustering Entire Parameters** in which all the parameters in all the layers are clustered into c number of clusters and a single table of centroids with c entries.
- 2) **Per-Layer Clustering of Parameters** in which the parameters of each individual layer are clustered separately. This means that we will have c number of clusters and a single table of centroids for each individual layer. Assuming l as the number of layers, we will have l separate tables of centroids each with c entries.

As Figure 6 (a) shows, parameters in different layers are all sharing the same table of centroids (highlighted in one single color), whereas when performing per-layer clustering, see Figure 6 (b), each layer has its own table of centroids (highlighted in various colors). Using the clustering techniques, the model size and therefore, memory footprint and memory transfers significantly reduce. In the baseline models, parameters are represented using FP32 (single-precision 32-bit floating-point). However, using the clustered parameters, each of the parameters can be represented with only 8 bits to index up to 256 distinct clusters. Although, in theory, less bits are needed to index less number of clusters, e.g., 6 bits for 64 clusters or 5 bits for 32 clusters, however, due to the complexity in alignment and handling data in these formats, in practice, they are rarely used. Therefore, in case of using less number of clusters than 256, the 8-bit index is still used for the sake of simplicity and data alignment in the memory.

In the following section, we present the results on various metrics when using clustering techniques on different transformers.

IV. METHODOLOGY AND EXPERIMENTAL SETUP

In this section, we present our methodology and experimental setup employed in this paper.

A. Platform and Hardware Setup

In order to demonstrate the highest gains of employing clustering schemes when supported in commonly-used platforms, we model three platforms with architectural characteristics similar to the following platforms for our experiments:

- 1) Conf-1: High-end Desktop Configuration. We modeled a desktop system featuring an NVIDIA-like GPU with 4352 CUDA cores and 11 GB GDDR6 memory, similar to a 2080 Ti, an Intel-like processor with 8 cores, and 64 GB of DDR4 Memory.
- 2) Conf-2: NVIDIA Tegra X2 System-on-Chip (SoC) [8]. Similarly, we model an NVIDIA TX2-like system featuring a quad-core Arm Cortex-A57 and a dual-core Denver CPU and a 256-core Pascal-based GPU.
- 3) Conf-3: NVIDIA AGX Xavier SoC [9] which features an octa-core Arm-based processor and a 512-core GPU.

Based on our analysis, an efficient hardware modification to support indirect access, which is the key element in implementing clustering schemes, can provide significant benefits in terms of performance improvements and energy consumption.

B. Datasets

For the evaluations and experimental results, accuracy, and performance analysis in this paper, we use the ImageNet [20] validation dataset.

C. Evaluation Metrics

In our evaluations, we consider the following metrics:

- *Speedup*. We show the obtained speedup when applying clustering.
- *Accuracy*. We compare the relative accuracy change with respect to the baseline model when applying clustering.
- *Memory Usage*. We compare the model size in Megabytes (MB) before and after applying the clustering. Model size reduction is equivalent to compression ratio of the model. With this metric we will demonstrate the usage of the memory bandwidth usage as well as the memory storage requirements to store the model parameters.
- *Energy Savings*. We compare the overall estimated energy consumption of the baseline compared to the cluster model.

D. Implementations

We use PyTorch [21] to run the models and measure the accuracy. We have implemented our own kernels to perform the clustering of the parameters. For each of the GPU platforms, we fine-tune the parameters to gain the best performance. We use our simulator to measure the performance and timings of the kernels using clustering. To measure the energy consumption accurately for the the platforms similar to NVIDIA TX2 and NVIDIA AGX Xavier platforms, we have accessed the integrated power, current, and voltage rails provided by

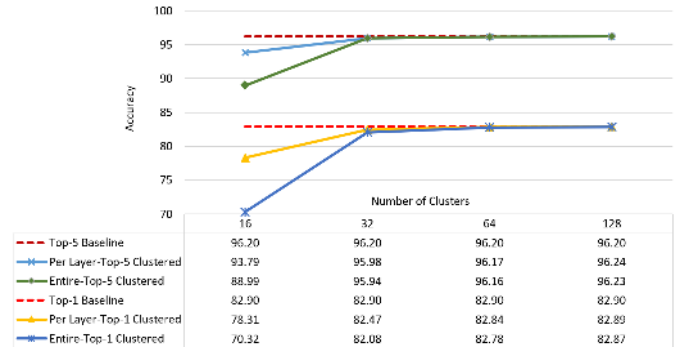


Fig. 7. Top-1 and Top-5 accuracy of the DeiT model when using clustering.

NVIDIA thermal guide on TX2 and Xavier [22], [23]. By periodically accessing these rails through specified registers and reading them, we calculate the energy consumption of each unit (e.g., DDR memory, GPU SoC, etc.) for the period of time in which the task is running. We use CACTI 6.5 [24] to model the energy consumption of the table of centroids.

V. EXPERIMENTAL RESULTS

In this section, we first present the results on how clustering techniques affect the accuracy of the different transformer models. Second, we present the performance and energy savings while employing the clustered models. Finally, we discuss the trade-offs between accuracy, memory requirements, performance improvements and energy savings.

A. Accuracy

We applied clustering techniques for the ViT and DeiT models. Figures 7 and 8 show the top-1 and top-5 accuracy results for different number of clusters for DeiT and ViT models respectively. Overall, top-1 accuracy results are lower than top-5 results [10], [15]. As Figure 7 shows, when using less number of clusters, e.g. 16 clusters, the per-layer clustering provides significantly higher accuracy. By increasing the number of clusters the accuracy results reach as good as in the baseline model. For instance, when using 64 clusters for the DeiT model, the top-1 and top-5 accuracy are only 0.1% and 0.05% lower than the baseline respectively, which is negligible. By using 128 or more number of clusters, there is zero accuracy loss.

As Figure 8 shows, we observe similar trend of results, as expected, for the ViT model. We can see that clustering performs the same for this model and by using only 64 clusters for the DeiT model, the top-1 and top-5 accuracy are only 0.3% and 0.2% lower than the baseline respectively, which is considered negligible. In these experiments we do not explore more than 128 number of cluster since they provide similar results as the baseline. Similarly, using less than 16 clusters, in most cases, it cannot capture the complexity of the model and results in very high accuracy loss. Based on our analysis and the related work discussed in Section VI, we conclude that clustering schemes work very well for transformers and they can offer significant benefits.

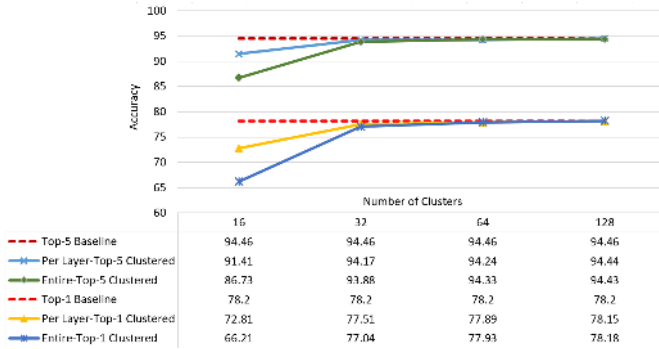


Fig. 8. Top-1 and Top-5 accuracy of the ViT model when using clustering.

B. Performance Improvements (Speedup)

To measure the maximum performance gain when using clustered parameters, we have developed and optimized a specific kernel to operate on clustered data. By fine-tuning and executing our kernel on each of the modeled GPUs, we have observed 5% to 38% speedup as shown in Figure 9. As it shows, despite extra instructions and overhead in the kernel to perform the indirect accesses, as shown in Figure 5, the reduced pressure on the memory system, because of clustered parameters, provides significant benefit specially in GPUs with more computing resources such as the GPU in Conf-3. To demonstrate the advantages of using clustered data, the results are obtained while putting maximum pressure on the memory subsystem. In particular, we have created controlled traffic on the memory subsystem to make limited bandwidth available for our experiments. This is done by concurrently running memory-intensive tasks putting pressure on the memory subsystem. Therefore it leaves less bandwidth for our experiments. As discussed earlier, in modern SoCs, multiple computing units need to access the shared memory.

Ideal Case. To demonstrate the maximum possible performance gains in an ideal system, we have considered a scenario in which the GPU computation power is fully underutilized due to lack of sufficient memory bandwidth. Assuming that the number of computation units is relatively larger than the memory capacity to feed them with data, the GPU can become underutilized. Although in modern GPUs this is a key design factor and it is considered while designing, however, in more powerful and specialized accelerators, this imbalance can become inevitable and therefore, result in performing far from peak capacity. According to Amdahl’s law [25], assuming to have enough computing resources, the reduction in the memory bandwidth and parameters size can increase the computation per memory unit ratio and provide significant performance gains as demonstrated in Figure 9. We have calculated the maximum speedup possible according to Amdahl’s law using an analytical model.

C. Memory Usage

When applying clustering on the model parameters, the 32-bit parameters are replaced by 8-bit index values. This results

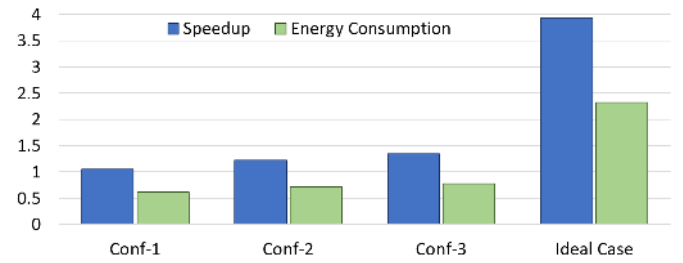


Fig. 9. Speedup and energy consumption normalized to the baseline implementation on different modeled platforms.

in a reduction of 4x in memory usage and memory bandwidth usage. Note that the table of centroids require a very small memory space relatively. For instance, for 64 clusters, the table of centroids occupies only 256 bytes.

D. Energy Savings

Figure 9 shows the normalized energy consumption in each of the configuration when using clustering. As it shows in the Conf-1, Conf-2, and Conf-3 the energy consumption reduces by 39%, 22%, and 22% respectively. The highest reduction belongs to the Conf-1 since the memory subsystem takes a considerable portion of the overall energy consumption. The memory subsystem can become the bulk of energy consumption in specialized accelerators [26]. In this case, the energy savings can even provide higher savings overall.

Ideal Case. Similar to the aforementioned discussion regarding the ideal performance, upon achieving the ideal performance, the energy consumption can be drastically reduced as Figure 9 shows. Note that speedup by means of increase in FLOPS usually translates into operating at a higher power. Despite this, due to a shorter execution time considerable static energy is saved in addition to the savings from the reduced memory traffic.

E. Discussion

The use of clustering model parameters can have numerous objectives and depending on the user’s requirements, several factors are considered.

Accuracy. Accuracy is the key factor for the user and depending on the task, various thresholds for accuracy loss can be considered in return for performance improvements or energy savings. In other words, depending on user’s demands, accuracy loss introduced by using clustering techniques is tolerable. In some cases, achieving the highest accuracy possible is the only target regardless of costs and resource requirements. However, in embedded domain and resource-constrained devices, achieving an acceptable level of performance and resource-usage is always considered. Therefore, such approaches are commonly employed with negligible, and in some cases, with zero accuracy loss.

Using Generalized Processors vs. Accelerators. General-purpose architectures or even GPUs are featuring flexible and programmable design to be used for variety of algorithms

and domains. Although it is trivial to implement clustering techniques for CPUs and GPUs, however, it can become costly since the implementations simply translates into more instructions to execute and irregular memory and cache access limiting the performance improvements and in some cases causing slowdown. Specialized accelerators, however, can be perfectly designed to obtain highest benefit from approaches such as clustering. Unlike general-purpose processors and GPUs, accelerators will be less flexible by design, however, they can highly perform complicated operations in a more efficient way, if they are designed to do so.

Applicability in Other Domains. In this paper, we focus on the transformers designed and trained for classification tasks, however, as we will discuss in Section VI, the use of clustering schemes and the obtained benefits can be generalized and extended to other domains such as segmentation [27], Language Modeling [28], other tasks in NLP, and any other domain that transformers are employed. Therefore, similar analysis, discussion, trade-offs, and conclusions can be drawn for those models. It is also proven that the accuracy and the impact of clustering techniques on other domains which are using other forms of deep learning models is similar to our study in this paper.

VI. RELATED WORK

Transformers. Ivanov et al. [4] state that data movement is the key bottleneck for training of transformers. The massive data requirements during the training, this process becomes memory-bound. Authors present a recipe for globally optimizing data movement in transformers. Their approach is able to reduce the data movement by more than 20% while achieving 30% speedup on BERT [3]. Although authors claim that their approach is beneficial for other forms of transformers, it is not clear to which extend those models get benefit. In this work, we perform a totally different approach which is orthogonal to the approach presented in [4]. Furthermore, our approach provides manifold performance and energy improvements while drastically reducing the memory bandwidth requirements.

Polino et al. [29] studied the effect of quantization which is mainly focusing on training. They proposed a technique called quantized distillation and leverage distillation during the training process, and a second technique, differentiable quantization, to optimizes the location of quantization points through stochastic gradient descent, to better fit the behavior of the model. Other works also performed similar studies and employed quantization during the training process or as a post-training step [30]–[32]. In this paper, we target the trained models and we proposed a different clustering scheme. Also, we focus on resource-constrained devices and provide extensive performance and energy analysis.

Shen et al. [33] studied different forms of quantization on deep bidirectional transformers BERT [3]. They performed an extensive analysis of fine-tuned BERT models using second order Hessian information in order to propose a novel method for quantizing BERT models to ultra low precision. In [19], authors quantized a trained Transformer machine language

translation model to lower precision 8-bit integers. They also reported performance analysis for Intel MKL Library. In [34], authors propose to use fixed-point arithmetic. The fixed-point optimization steps consist of quantization sensitivity analysis, hardware conscious word-length assignment, quantization and retraining, and post-training for improved generalization.

Other Models. Han et al. [35] performed model pruning and post-training clustering, and Huffman coding to compress multiple computer vision models. They further proposed an specialized accelerator [7] to better explore the sparsity and the clustered parameters and achieved four orders of magnitude energy reduction over the baseline models running on a CPU. Tabani et al. [26], [36] designed an accelerator for acoustic scoring process in hybrid large-vocabulary speech recognition systems. They show that despite orders of magnitude performance and energy consumption improvements using their design, clustering the parameters provides additional manifold speedup and energy savings. These works target different models such as CNNs or other domains such as speech recognition which are very different than transformers.

Similar to previous deep learning models, quantization schemes are well studied for transformers, including for both training and post-training processes, however, we believe that studies remained focused on commonly-used schemes such as vector and scalar quantization rather than techniques such as clustering of the data. Furthermore, the main focus of the previous work is on the accuracy and reasoning behind accuracy change when applying those techniques while in this work, we devote an important part of the paper to performance-related aspect. We also consider variety of frameworks and heterogeneous platforms which, to our knowledge, has not been explored earlier.

VII. CONCLUSIONS

In this paper, we first performed an extensive analysis of state-of-the-art vision transformers. We showed that applying clustering techniques on these models can provide significant speedup and energy savings on various platforms while having negligible drop in the overall accuracy. We presented the existing trade-offs and challenges towards improving the performance and energy savings and how it can be generalized to transformers trained for other domains. Our results on representative platforms show that significant energy savings is achievable while highest speedups can be reached in specialized accelerators.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
- [2] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *arXiv preprint arXiv:2101.01169*, 2021.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [4] A. Ivanov, N. Dryden, T. Ben-Nun, S. Li, and T. Hoefler, "Data movement is all you need: A case study on optimizing transformers," *arXiv e-prints*, pp. arXiv-2007, 2020.
- [5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.
- [6] J. Khan, "How next-gen AI accelerators will transform mobile machine learning: A look at how AI-dedicated hardware promises to bring end-to-end machine learning to smartphones," <https://heartbeat.fritz.ai/how-next-gen-ai-accelerators-will-transform-mobile-machine-learning-47d262dbf5d5>
- [7] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "Eie: Efficient inference engine on compressed deep neural network," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 243–254, 2016.
- [8] "NVIDIA - Jetson TX2 Module," <https://developer.nvidia.com/embedded/jetson-tx2>.
- [9] D. Shapiro, "Introducing Xavier, the NVIDIA AI Supercomputer for the Future of Autonomous Transportation," *NVIDIA blog*, 2016. [Online]. Available: <https://blogs.nvidia.com/blog/2016/09/28/xavier/>
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.
- [11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.
- [12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
- [14] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *arXiv preprint arXiv:2009.06732*, 2020.
- [15] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers and distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020.
- [16] G. P. Perrucci, F. H. Fitzek, and J. Widmer, "Survey on energy consumption entities on the smartphone platform," in *2011 IEEE 73rd vehicular technology conference (VTC Spring)*. IEEE, 2011, pp. 1–6.
- [17] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.
- [18] O. Zafrir, G. Boudoukh, P. Izsak, and M. Wasserblat, "Q8bert: Quantized 8bit bert," *arXiv preprint arXiv:1910.06188*, 2019.
- [19] A. Bhandare, V. Sripathi, D. Karkada, V. Menon, S. Choi, K. Datta, and V. Saletore, "Efficient 8-bit quantization of transformer neural machine language translation model," *arXiv preprint arXiv:1906.00532*, 2019.
- [20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: <http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [22] NVIDIA, "NVIDIA - Jetson TX2 Thermal Design Guide," 2017.
- [23] "NVIDIA - Jetson AGX Xavier Series Thermal Design Guide," 2020.
- [24] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi, "Cacti 5.1," Technical Report HPL-2008-20, HP Labs, Tech. Rep., 2008.
- [25] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proceedings of the April 18-20, 1967, spring joint computer conference*, 1967, pp. 483–485.
- [26] H. Tabani, J. Arnau, J. Tubella, and A. González, "An ultra low-power hardware accelerator for acoustic scoring in speech recognition," in *2017 26th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 2017, pp. 41–52.
- [27] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," *arXiv preprint arXiv:2012.15840*, 2020.
- [28] C. Wang, M. Li, and A. J. Smola, "Language models with transformers," *arXiv preprint arXiv:1904.09408*, 2019.
- [29] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," *arXiv preprint arXiv:1802.05668*, 2018.
- [30] A. Fan, P. Stock, B. Graham, E. Grave, R. Gribonval, H. Jegou, and A. Joulin, "Training with quantization noise for extreme fixed-point compression," *arXiv preprint arXiv:2004.07320*, 2020.
- [31] H. Wu, P. Judd, X. Zhang, M. Isaev, and P. Micikevicius, "Integer quantization for deep learning inference: Principles and empirical evaluation," *arXiv preprint arXiv:2004.09602*, 2020.
- [32] H. Bai, W. Zhang, L. Hou, L. Shang, J. Jin, X. Jiang, Q. Liu, M. Lyu, and I. King, "Binarybert: Pushing the limit of bert quantization," *arXiv preprint arXiv:2012.15701*, 2020.
- [33] S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "Q-bert: Hessian based ultra low precision quantization of bert," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8815–8821.
- [34] Y. Boo and W. Sung, "Fixed-point optimization of transformer neural network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1753–1757.
- [35] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.
- [36] H. Tabani, "Low-power architectures for automatic speech recognition," 2018.