T1000: Mitigating the memory footprint of convolution neural networks with decomposition and re-fusion

Changxi Liu a, Hailong Yang a, Rui Wang a,*, Zhongzhi Luan a, Depei Qian a,b

a Sino-German Joint Software Institute, School of Computer Science and Engineering, Beihang University, Beijing, 100191, China
b School of Data and Computer Science, Sun Yat-sen University, Guangzhou, 510275, China

**HIGHLIGHTS**

- We identify the memory problem when applying CP-decomposition to CNNs.
- We propose a decomposition and re-fusion approach to mitigate the memory problem.
- We demonstrate the effectiveness of our approach on two state-of-the-art CNNs.
- Our experiments with AlexNet and VGG-19 show great memory reduction and speedup.

**ABSTRACT**

In recent years, convolution neural networks have significantly advanced the frontier of computer vision and other intelligent applications due to its promising accuracy. However, the improved accuracy comes with the formidable computation complexity with deeper convolution layers, which prevents its adoption on resource constrained system such as embedded and mobile. Although research efforts have been devoted to reduce the computation complexity of convolution neural networks through tensor decomposition, the volume of intermediate data generated by the tensor decomposition grows dramatically, which consumes more memory resource and has not been addressed by existing work.

In this work, we propose T1000 to re-fuse the convolutions across tensors after applying the canonical polyadic decomposition to conventional convolution layers so that we can receive the benefit of reduced computation complexity, in the meanwhile mitigate the memory occupancy of the intermediate data. We demonstrate the effectiveness of our approach by applying canonical polyadic decomposition and re-fusion to the convolution layers of two well-known convolution neural networks, AlexNet and VGG-19 implemented with Caffe. Compared to the default canonical polyadic decomposition, our approach reduces the memory occupancy of the intermediate data by 84.6% and 77.4% for AlexNet and VGG-19 respectively. In addition, our approach improves the performance of AlexNet and VGG-19 by 1.77 × and 1.4 × respectively.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

In the recent years, the advancement in image classification [1,2] and object detection [3,4] achieved by convolution neural networks (CNNs) have demonstrated that deep learning is an effective approach to develop intelligent computer vision applications, such as self-driving car, personal assistant and artificial intelligent robot. However, as the depth of neural network grows, the computation demand of CNNs is becoming a major obstacle preventing its pervasive adoption. Even though the training tasks can be done on high-end server with powerful accelerators (e.g., GPU and FPGA), there are rising interests from both industry and academia to deploy inference tasks in resource constrained fields such as embedded system and mobile device [5–7]. Due to the limited computation and memory capacity, it is critical to mitigate the resource consumption of CNNs for its successful adoption in resource constrained fields.

To address the above challenges from different perspectives, there has been growing amount of research works such as developing smaller networks with negligible precision loss [8–11], advancing the mathematical computation method [12,13] and modifying existing networks to adapt to the hardware architecture [5,14–16].
(e.g., transforming the convolution layers to reduce computation complexity). Among these studies, the idea of convolution dimensionality reduction (e.g., tensor decomposition [17]) is considered to be an effective way to mitigate the computation complexity. However, existing work [18,5,16] fails to consider the intermediate data generated after the transformation, which consumes significant amount of memory resource.

Among the convolution dimensionality reduction approach [18, 17,5,16], canonical polyadic decomposition (CP-decomposition) [5] has been widely used by researchers to optimize the convolution operations. The CP-decomposition actually involves two steps such as kernel decomposition and convolution composition. The decomposition means breaking the high-dimensional convolution kernel tensors into a series of low rank ones. Whereas the composition means replacing the original convolution layer with a composition of four convolution layers with decomposed kernel tensors. The fundamental idea is similar to service composition in the field of cloud computing [19,20]. By approximating a multidimensional convolution to the sum of several low-rank tensors, CP-decomposition can effectively cut down the number of convolution operations with negligible precision loss (detailed discussion in Section 2.2). However, when applying CP-decomposition to convolution layers, it generates large amount of intermediary data by low-rank tensors, exacerbating the problem of memory consumption. To further illustrate, we apply CP-decomposition to the most time consuming convolution layers (e.g., conv2) of AlexNet [21] and VGG-19 [22], and measure memory footprint of each convolution layer after CP-decomposition. Note that the precision loss of both networks is less than 1% after applying the CP-decomposition. The experimental details are shown in Section 4.1.

As shown in Fig. 1, the left two bars represent the results of AlexNet, while the right two bars represent the results of VGG-19. The memory footprint of AlexNet and VGG-19 is shown on the left y-axis and right y-axis respectively. The bar labeled Traditional-Conv shows the results of the original convolution layers, while the bar with CP-Conv shows the results after applying CP-decomposition. Comparing to the original convolution layer, CP-decomposition generates large amount of intermediate data while reducing the computational complexity. For instance, the size of intermediate data for AlexNet and VGG-19 increases by more than 26 x and 7 x respectively. The reason for the dramatic increase of intermediate data size is that CP-decomposition utilizes three cascaded tensors (much smaller than the original convolution kernels) to reduce the number of convolution operations. The intermediate data generated by previous tensor is passed to the next tensor, which requires additional memory space to store. Moreover, we observe that the volume of the intermediate data is closely related to the batchsize chosen by the network. Therefore, it is critical to mitigate the memory footprint of intermediary data from CP-decomposition for its successful adoption in resource constrained fields.

The idea of CNN layer re-fusion is explored by [14]. However, there are several challenges to be addressed in order to re-fuse the convolutions across tensors from CP-decomposition. (1) Since the original convolution is decomposed across several tensors, it remains unclear which convolutions should be re-fused and how the decision affects the memory occupancy as well as accuracy. (2) Convolution re-fusion itself consumes memory to store temporary data, therefore it is important to optimize the memory usage during re-fusion. (3) The intermediary data generated from one tensor is the input for the subsequent tensor. The efficiency of the convolution operations across tensors determine the performance of the layer after applying re-fusion.

To address the above challenges, we propose a decomposition and re-fusion approach T1000. It leverages the advantage of CP-decomposition to reduce the computation complexity of convolution operations, and then use re-fusion to mitigate the volume of intermediate data introduced by CP-decomposition. We evaluate the effectiveness of T1000 from several aspects by applying CP-decomposition and re-fusion to two state-of-the-art CNNs (AlexNet and VGG-19). The experiments results demonstrate that T1000 can effectively reduce the size of intermediary data while improving the performance of the original CNN models. Specifically, this paper makes the following contributions:

- We identify the memory problem due to the large amount of intermediary data introduced by CP-decomposition when it is applied to CNNs with comprehensive analysis.
- To overcome the memory problem, we propose a decomposition and re-fusion approach (T1000) to mitigate the volume of intermediary data through combining separate convolution operations across tensors into an integrated computation process.
- We demonstrate the effectiveness of T1000 on two state-of-the-art CNNs (AlexNet and VGG-19) that significantly reduces the size of intermediary data as well as improves the performance of convolution layers.

The rest of the paper is organized as follows. Section 2 describes the background of CNNs and CP-decomposition that motivates our study. Section 3 proposes our decomposition and re-fusion approach. Section 4 details the evaluation. Section 5 discusses the related work. Section 6 concludes our work.

2. Background

2.1. Convolution neural networks

Neural networks have been demonstrated success in many fields such as speech signal prediction [23] and medical data classification [24]. Especially the Convolution Neural Network (CNN) [21] have received dramatic research attention recently with its extraordinary performance on ImageNet [25]. CNN is a multiple-layer neural network in which the connectivity pattern between its neurons is inspired by the organization of the animal visual cortex. In general, CNNs is composed by series of convolution and pooling layers followed by classification layer. The main function of convolution operation is to extract features from input, and then produce high level feature maps as the layer goes up. Taking advantage of the local perception and parameter sharing, CNNs is capable of reducing the massive weight parameters compared to the fully-connected neural network. The feature maps are computed through convolution between feature maps and convolution kernels. When a three dimensional feature map with depth M performs convolution with N convolution kernels of the
دریافت فوری متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات