arXiv:2412.01199v1 [cs.CV] 2 Dec 2024

TinyFusion: Diffusion Transformers Learned Shallow

Gongfan Fang*, Kunjun Li*, Xinyin Ma, Xinchao Wang†
National University of Singapore
{gongfan, kunjun, maxinyin}@u.nus.edu, xinchao@nus.edu.sg

Abstract

Diffusion Transformers have demonstrated remarkable capabilities in image generation but often come with excessive parameterization, resulting in considerable inference overhead in real-world applications. In this work, we present TinyFusion, a depth pruning method designed to remove redundant layers from diffusion transformers via end-to-end learning. The core principle of our approach is to create a pruned model with high recoverability*, allowing it to regain strong performance after fine-tuning. To accomplish this, we introduce a differentiable sampling technique to make pruning learnable, paired with a co-optimized parameter to simulate future fine-tuning. While prior works focus on minimizing loss or error after pruning, our method explicitly models and optimizes the post-fine-tuning performance of pruned models. Experimental results indicate that this learnable paradigm offers substantial benefits for layer pruning of diffusion transformers, surpassing existing importance-based and error-based methods. Additionally, TinyFusion exhibits strong generalization across diverse architectures, such as DiTs, MARs, and SiTs. Experiments with DiT-XL show that TinyFusion can craft a shallow diffusion transformer at less than 7% of the pre-training cost, achieving a 2× speedup with an FID score of 2.86, outperforming competitors with comparable efficiency. Code is available at* `https://github.com/VainF/TinyFusion`

1. Introduction

Diffusion Transformers have emerged as a cornerstone architecture for generative tasks, achieving notable success in areas such as image [11, 26, 40] and video synthesis [25, 59]. This success has also led to the widespread availability of high-quality pre-trained models on the Internet, greatly accelerating and supporting the development of various downstream applications [5, 16, 53, 55]. However, pre-trained diffusion transformers usually come with considerable inference costs due to the huge parameter scale, which poses significant challenges for deployment. To resolve this problem, there has been growing interest from both the research community and industry in developing lightweight models [12, 23, 32, 58].

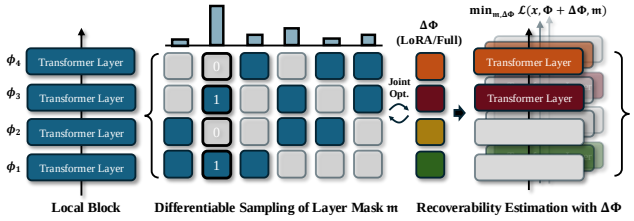

Figure 1. This work presents a learnable approach for pruning the depth of pre-trained diffusion transformers. Our method simultaneously optimizes a differentiable sampling process of layer masks and a weight update to identify a highly recoverable solution, ensuring that the pruned model maintains competitive performance after fine-tuning.

The efficiency of diffusion models is typically influenced by various factors, including the number of sampling steps [33, 43, 45, 46], operator design [7, 48, 52], computational precision [19, 30, 44], network width [3, 12] and depth [6, 23, 36]. In this work, we focus on model compression through depth pruning [36, 54], which removes entire layers from the network to reduce the latency. Depth pruning offers a significant advantage in practice: it can achieve a linear acceleration ratio relative to the compression rate on both parallel and non-parallel devices. For example, as will be demonstrated in this work, while 50% width pruning [12] only yields a 1.6× speedup, pruning 50% of the layers results in a 2× speedup. This makes depth pruning a flexible and practical method for model compression.

This work follows a standard depth pruning framework: unimportant layers are first removed, and the pruned model is then fine-tuned for performance recovery. In the literature, depth pruning techniques designed for diffusion transformers or general transformers primarily focus on heuristic approaches, such as carefully designed importance scores [6, 36] or manually configured pruning

*Equal contribution
†Corresponding author