Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention

Angelos Katharopoulos [1 2] **Apoorv Vyas** [1 2] **Nikolaos Pappas** [3] **François Fleuret** [2 4 *]

Abstract

Transformers achieve remarkable performance in several tasks but due to their quadratic complexity, with respect to the input's length, they are prohibitively slow for very long sequences. To address this limitation, we express the self-attention as a linear dot-product of kernel feature maps and make use of the associativity property of matrix products to reduce the complexity from $\mathcal{O}\left(N^2\right)$ to $\mathcal{O}\left(N\right)$, where N is the sequence length. We show that this formulation permits an iterative implementation that dramatically accelerates autoregressive transformers and reveals their relationship to recurrent neural networks. Our *linear transformers* achieve similar performance to vanilla transformers and they are up to 4000x faster on autoregressive prediction of very long sequences.

1. Introduction

Transformer models were originally introduced by Vaswani et al. (2017) in the context of neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2015) and have demonstrated impressive results on a variety of tasks dealing with natural language (Devlin et al., 2019), audio (Sperber et al., 2018), and images (Parmar et al., 2019). Apart from tasks with ample supervision, transformers are also effective in transferring knowledge to tasks with limited or no supervision when they are pretrained with autoregressive (Radford et al., 2018; 2019) or masked language modeling objectives (Devlin et al., 2019; Yang et al., 2019; Song et al., 2019; Liu et al., 2020).

However, these benefits often come with a very high computational and memory cost. The bottleneck is mainly caused by the global receptive field of self-attention, which processes contexts of N inputs with a quadratic memory and time complexity $\mathcal{O}\left(N^2\right)$. As a result, in practice transformers are slow to train and their context is *limited*. This disrupts temporal coherence and hinders the capturing of long-term dependencies. Dai et al. (2019) addressed the latter by attending to memories from previous contexts albeit at the expense of computational efficiency.

Lately, researchers shifted their attention to approaches that increase the context length without sacrificing efficiency. Towards this end, Child et al. (2019) introduced sparse factorizations of the attention matrix to reduce the self-attention complexity to $\mathcal{O}\left(N\sqrt{N}\right)$. Kitaev et al. (2020) further reduced the complexity to $\mathcal{O}\left(N \log N\right)$ using locality-sensitive hashing. This made scaling to long sequences possible. Even though the aforementioned models can be efficiently trained on large sequences, they do not speed-up autoregressive inference.

In this paper, we introduce the *linear transformer* model that significantly reduces the memory footprint and scales linearly with respect to the context length. We achieve this by using a kernel-based formulation of self-attention and the associative property of matrix products to calculate the self-attention weights (§ 3.2). Using our linear formulation, we also express causal masking with linear complexity and constant memory (§ 3.3). This reveals the relation between transformers and RNNs, which enables us to perform autoregressive inference orders of magnitude faster (§ 3.4).

Our evaluation on image generation and automatic speech recognition demonstrates that *linear transformer* can reach the performance levels of transformer, while being up to three orders of magnitude faster during inference.

2. Related Work

In this section, we provide an overview of the most relevant works that seek to address the large memory and computational requirements of transformers. Furthermore, we discuss methods that theoretically analyze the core component of the transformer model, namely self-attention. Finally, we present another line of work that seeks to alleviate the softmax bottleneck in the attention computation.

[1]Idiap Research Institute, Switzerland [2]EPFL, Switzerland [3]University of Washington, Seattle, USA [4]University of Geneva, Switzerland. [*]Work done at Idiap. Correspondence to: Angelos Katharopoulos <firstname.lastname@idiap.ch>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.