TCNCA: Temporal Convolution Network with Chunked Attention for Scalable Sequence Processing

Aleksandar Terzić[1,2*], **Michael Hersche**[1,2], **Geethan Karunaratne**[1]
Luca Benini[2], **Abu Sebastian**[1], **Abbas Rahimi**[1†]
[1]IBM Research – Zurich, [2]ETH Zurich

Abstract

MEGA is a recent transformer-based architecture, which utilizes a linear recurrent operator whose parallel computation, based on the FFT, scales as $O(LlogL)$, with L being the sequence length. We build upon their approach by replacing the linear recurrence with a special temporal convolutional network which permits larger receptive field size with shallower networks, and reduces the computational complexity to $O(L)$. The resulting model is called **TCNCA**, a **T**emporal **C**onvolutional **N**etwork with **C**hunked **A**ttention. We evaluate TCNCA on EnWik8 language modeling, long-range-arena (LRA) sequence classification, as well as a synthetic reasoning benchmark associative recall. On EnWik8, TCNCA outperforms MEGA, reaching a lower loss with $1.37\times$/$1.24\times$ faster forward/backward pass during training. The dilated convolutions used in TCNCA are consistently and significantly faster operations than the FFT-based parallelized recurrence in GPUs, making them a scalable candidate for handling very large sequence lengths: they are up to $7.07\times$/$2.86\times$ faster in the forward/backward pass for sequences up to 131 k. Further on LRA, TCNCA achieves, on average, $1.28\times$ speed-up during inference with similar accuracy to what MEGA achieves. On associative recall, we find that even a simplified version of TCNCA, without excessive multiplicative and additive interactions, remains superior or competitive to MEGA on a range of sequence lengths and vocabulary sizes.

1 Introduction

The Transformer [1] is a powerful class of neural networks which has found success in a variety of tasks including image processing [2], physical system modeling [3], drug discovery [4], but perhaps most notably, language modeling [5], [6], [7]. While undeniably a strong candidate for a universally applicable neural network, the operator at its backbone, *attention*, faces some crucial limitations. We consider two limitations, including the $O(L^2)$ computational and memory complexity [8] of attention, as well as its poor performance in long sequence classification, namely on the long-range-arena (LRA) dataset [9], where it is drastically outperformed by *linear recurrent models* [10–12]; however, these models lag behind the transformer on language modeling [13]. A more extensive review of related works can be found in Appendix A.

A recent neural network, MEGA [14], combines the strengths of *linear recurrences* and *attention* in a manner which scales sub-quadratically. Concretely, MEGA combines the damped exponential moving average (EMA) known from time-series analysis [15], with chunked attention which operates on fixed-size non-overlapping blocks in the input sequence. It achieves scores competitive with the state-of-the-art in a range of disparate tasks including language modeling on the EnWik8 dataset [16] and LRA sequence classification [9].

*Research conducted at IBM Research – Zurich.
†Corresponding author: abr@zurich.ibm.com

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

We focus on EMA, which maps $\mathbf{x_t} \in \mathbb{R}^h$ to $\mathbf{y_t} \in \mathbb{R}^h$ using the parameters $\alpha, \delta \in [0,1]^h, h \in \mathbb{N}_+$ as:

$$\mathbf{y_t} = \alpha \odot \mathbf{x_t} + (\mathbf{1} - \alpha \odot \delta) \odot \mathbf{y_{t-1}}. \tag{1}$$

This operation can be directly computed as per equation 1. However, during training and non-causal data processing, it can equivalently be computed as a convolution with a kernel which is of the same shape as the input data [14]. This convolution can be efficiently performed in $O(LlogL)$ time in the frequency domain [17], [10]. This mode of operation is interesting because it allows for a higher utilization of GPUs' parallel processing capabilities [17].

In this work, we investigate the performance and runtime effects of replacing the bottleneck EMA within the MEGA processing stack with a dedicated temporal convolutional neural network (TCN) [18–21], an operator which scales linearly with the sequence length. The TCN employs dilated convolutions, which allow the network to achieve a large receptive field with few parameters. TCNs are typically implemented as a cascade of *residual blocks*, in which each block applies two dilated convolution operations with equal dilations. In order to quickly reach large receptive fields, the dilation exponentially increases with each successive block [18, 21]. Our model differs from what is usually used in literature in that it only includes a single dilated convolution operation per residual block. This construction allows for a larger receptive field size with shallower networks. Details are given in Appendix E. We call the resulting model, which combines a TCN with chunked attention, TCNCA.

We find that on EnWik8 language modeling, TCNCA outperforms MEGA [14] (and Transformer-XL [22]), achieving a BPC score of 1.01, in addition to $1.37\times/1.24\times$ faster forward/backward pass. On a synthetic reasoning benchmark, *associative recall*, a simplified version of TCNCA (see Appendix C) is competitive with MEGA over a range of different sequence lengths and vocabulary sizes. On 64-dimensional sequences of lengths ranging from 8192 to 131072, the employed dilated convolution operator is up to $7.07\times$ and $2.86\times$ faster than the parallelized EMA of MEGA in the forward and backward pass, respectively. This signifies the scalability of the approach to long sequences thanks to its linear complexity. On the LRA classification tasks, TCNCA slightly underperforms MEGA by only 0.1% on average, while achieving $1.28\times$ inference speedup.

2 The TCNCA model

An overview of the model and the operations used therein is shown in Figure 1. At a high-level, the model can be thought of as a concatenation of a temporal convolutional neural network (Figure 1b) with chunked attention (Figure 1d). The sketch is simplified; the actual construction follows the one defined by MEGA [14], and is outlined in Appendix C.

Figure 1a shows a depth-N sequence processing stack. Each of the N-many layers consists of a temporal convolutional network and chunked attention, both of which operate along the time axis,

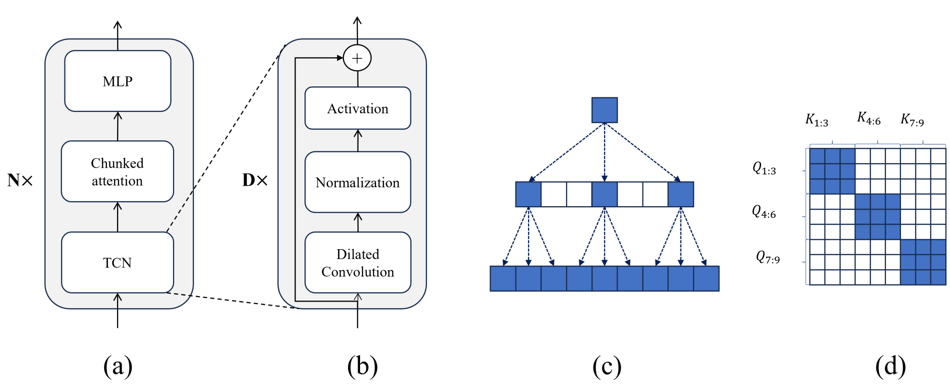


Figure 1: (a) Simplified high-level overview of the TCNCA model. (b) The TCN residual block. (c) Connectivity of a TCN with kernel size $K = 3$, dilation factor $f = 3$, and depth $D = 2$. (d) Chunked attention operation which computes query-key similarities in fixed-width non-overlapping windows, shown with chunk size 3.

followed by a multi-layer perceptron (MLP) operating along the feature axis. For each embedding dimension, a TCN with its own set of trainable parameters is instantiated.

The TCN block in Figure 1a is expanded in Figure 1b. Three integer hyperparameters govern the TCN construction; kernel size K, dilation factor f, and depth D. The TCN consists of D-many residual blocks, each of which implements a dilated convolution operation whose dilation is determined by the layer index $i = 0, ..., D-1$ and f as f^i. In Figure 1c, we show the connectivity pattern of a TCN with $D = 2$, $f = 3$ and $K = 3$.

Following the TCN, which scales as $O(L)$, we have chunked attention. As already noted, it computes the query-key similarities only within fixed-size non-overlapping windows within the sequence, as shown in Figure 1d. This is also an $O(L)$ operation.

3 Experiments

EnWik8 language modeling EnWik8 is a dataset which comprises a subset of the English Wikipedia. We train and evaluate our model on EnWik8 character-level language modeling in the same manner as was done in MEGA [14]. The results are shown in Table 1. More details are given in Appendix F.

Table 1: EnWik8 bit-per-character scores. Results marked with a star (*) are taken from [14].

Model	Transformer-XL	MEGA	TCNCA
BPC	1.06*	1.02*	**1.01**
Parameters	41M	39M	39M

TCNCA outperforms the Transformer-XL [22] as well as MEGA [14], reaching a 1.01 BPC score. For transparency's sake, we have to note that the scores reported in relevant literature are rounded down to 2 digits after the decimal point, hence we do the same. With 4 digits after the decimal point, the score we achieve is 1.0144 BPC.

We measure the forward and backward pass speed-up on a 16GB Nvidia V100 GPU during training. During training, TCNCA achieves a $\mathbf{1.373}\times$ speed-up in the forward pass and a $\mathbf{1.245}\times$ speed-up in the backward pass, compared to MEGA. However, speeding up the inference runtime of the generative tasks is not straightforward and is one of the limitations of this work (see Appendix B).

Long-range-arena Long-range-arena [9] comprises six classification tasks with sequence lengths ranging from 1024 to 16384. The benchmarks are varied, including pattern detection, sentiment classification, mathematical reasoning, and visual logical reasoning. We use the same dimensionalities, hyperparameters, and attention chunk sizes as those used in MEGA [14], and select the TCN construction as per Appendix D. Results are shown in Table 2.

Although TCNCA lags behind the state-of-the-art state space method, S5 [12], by 2.3%, it is on par with MEGA-chunk (just an average of a 0.1% lower accuracy) while achieving an average inference speed-up 28%.

Table 2: Long-range-arena accuracies (%) of state-of-the-art models. The Transformer scores are taken from the reproduction in MEGA [14]. All other results, excluding TCNCA, were taken from the respective papers. The last row reports the end-to-end inference speed-up of TCNCA measured against MEGA-chunk.

Model	ListOps	Text	Retrieval	Image	Path	Path-X	Average
Transformer [1] [14]	37.1	65.2	79.1	42.9	71.8	50	57.7
S4D [23]	60.5	86.2	89.5	89.9	93.1	91.9	85.2
S5 [12]	62.2	89.3	91.4	90.1	95.3	98.6	87.8
LRU [11]	60.2	89.4	89.9	89.0	95.7	96.0	86.7
SGConv [24]	61.4	89.2	91.1	87.97	95.4	97.8	87.1
MEGA chunk [14]	58.7	90.2	91.0	85.8	94.4	93.8	85.6
TCNCA	59.6	89.8	89.4	86.8	94.5	92.7	85.5
Speedup (forward pass)	1.05×	1.25×	1.18×	1.24×	1.25×	1.73×	1.28×

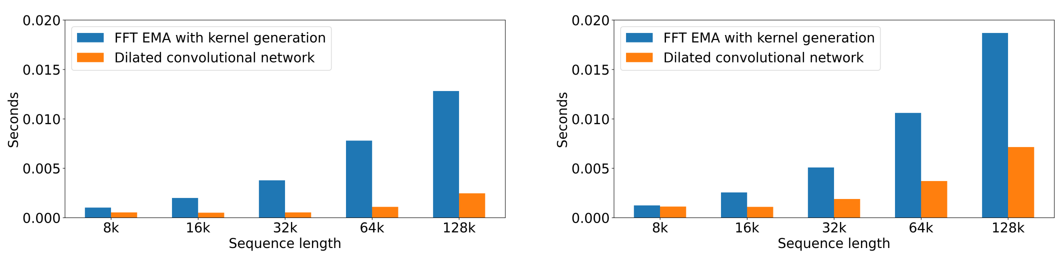
Table 3: Associative recall accuracy (%) with varying sequence lengths and vocabulary sizes.

	Vocabulary size 10		Vocabulary size 20	
Seq. len.	MEGA	TCNCA-simple	MEGA	TCNCA-simple
64	98.8	100	62.4	56
1024	99.6	100	99.4	97.6
4096	100	100	100	99.6
8192	98.2	100	98.6	99.2

Associative recall This synthetic benchmark requires faithful attention and measures the basic reasoning capability of neural sequence models, remembering associations between pairs of tokens [25] [13]. For example, given a sequence of tokens *a 2 c 4 b 3 d 1*, if the model is prompted with *a*, the expected output is *2*, the token following *a* in the input sequence. If it were prompted with *b*, the correct output would be *3*, etc.

As mentioned, TCNCA is based on MEGA [14], and as such it involves an intricate interconnection between the different modules it is composed of. We report TCNCA scores for the associative recall in a setting in which the module interconnection is significantly simplified by eliminating excessive multiplicative and additive interactions (TCNCA-simple, see Appx. C). Over the investigated range of vocabulary sizes and sequence lengths in Table 3, TCNCA-simple remains competitive with MEGA.

Parallelized EMA vs. dilated convolution runtime measurements We measure the forward and backward-pass runtimes of a dilated convolutional network and a parallelized EMA recurrence over a range of sequence lengths, and report the results in Figure 2. For a clear comparison of the two operations, we strip both of them of residual connections, non-linearities as well as normalization layers. They are roughly parameter-matched, with EMA having 64 parameters and the dilated convolution having 68 parameters. The dilated convolutional network is configured with $K = 17$, $D = 4$, and f is increased until the receptive field of the network is larger than the sequence length it operates on. The benchmarks were run on an Nvidia V100 with 16 GB of VRAM. Further details are given in Appendix H.



(a) Forward pass runtime measurements.

(b) Backward pass runtime measurements.

Figure 2: Run-time comparisons between a parallel linear recurrence including kernel generation (blue) and a dilated CNN (orange) for the forward and backward pass, with varying sequence lengths. The dilated convolutional network is consistently the faster operation.

4 Conclusion

In this work inspired by ground-breaking results from the team behind MEGA [14], we show that a TCN and chunked attention hybrid model, TCNCA, is able to compete with the state-of-the-art models on Enwik8 language modeling and Long-Range-Arena sequence classification. During training and non-causal inference workloads, TCNCA consistently exhibits inference speed-ups in the range of 5% to 73% compared to MEGA-chunk. We show that a simplified version of TCNCA solves the *associative recall* synthetic reasoning benchmark with a similar accuracy as does MEGA. Finally, we show that on the Nvidia V100 GPU, a dilated convolutional network is consistently faster than an FFT-based parallelized EMA recurrence over a wide range of sequence lengths. Some of the limitations of our approach are detailed in Appendix B.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021.

[3] Nicholas Geneva and Nicholas Zabaras. Transformers for modeling physical systems. *Neural Networks*, 146:272–289, 2022.

[4] Daria Grechishnikova. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Scientific reports*, 11(1):321, 2021.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (ACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 1877–1901, 2020.

[8] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55:1 – 28, 2020.

[9] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations (ICLR)*, 2021.

[10] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

[11] Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning (ICML)*, 2023.

[12] Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

[13] Tri Dao, Daniel Y. Fu, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry Hungry Hippos: Towards language modeling with state space models. In *International Conference on Learning Representations (ICLR)*, 2023.

[14] Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Zettlemoyer Luke. Mega: Moving average equipped gated attention. In *International Conference on Learning Representations (ICLR)*, 2023.

[15] Eddie McKenzie and Everette S. Gardner. Damped trend exponential smoothing: A modelling viewpoint. *International Journal of Forecasting*, 26(4):661–665, 2010.

[16] Marcus Hutter. The human knowledge compression contest, 2006.

[17] Narsimha Reddy Chilkuri and Chris Eliasmith. Parallelizing legendre memory unit training. In *International Conference on Machine Learning (ICML)*, pages 1898–1907, 2021.

[18] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*, 2018.

[19] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016.

[20] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[21] Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. Eeg-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2958–2965, 2020.

[22] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2978–2988, 2019.

[23] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:35971–35983, 2022.

[24] Yuhong Li, Tianle Cai, Yi Zhang, Deming Chen, and Debadeepta Dey. What makes convolutional models great on long sequence modeling? In *International Conference on Learning Representations (ICLR)*, 2023.

[25] Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.