Replacing softmax with ReLU in Vision Transformers

Mitchell Wortsman Jaehoon Lee Justin Gilmer Simon Kornblith
Google DeepMind

Abstract

Previous research observed accuracy degradation when replacing the attention softmax with a point-wise activation such as ReLU. In the context of vision transformers, we find that this degradation is mitigated when dividing by sequence length. Our experiments training small to large vision transformers on ImageNet-21k indicate that ReLU-attention can approach or match the performance of softmax-attention in terms of scaling behavior as a function of compute.

1 Introduction

The transformer architecture [24] is ubiquitous in modern machine learning. Attention, a central component of the transformer [2], includes a softmax which produces a probability distribution over tokens. Softmax is costly due to an exponent calculation and a sum over sequence length which makes parallelization challenging [22, 6].

In this report we explore point-wise alternatives to the softmax operation which do not necessarily output a probability distribution. As a highlight, we observe that attention with ReLU divided by sequence length can approach or match traditional softmax attention in terms of scaling behavior as a function of compute for vision transformers. This result presents new opportunities for parallelization, as ReLU-attention can be parallelized over the sequence length dimension with fewer gather operations than traditional attention.

2 Related work

Previous research has explored substituting softmax with ReLU [23, 12] or squared ReLU [13]. However, these approaches do not divide by sequence length, which we experimentally find is important to reach accuracy comparable to softmax. In addition, previous research [19] has replaced softmax while still requiring normalization over the sequence length axis to ensure

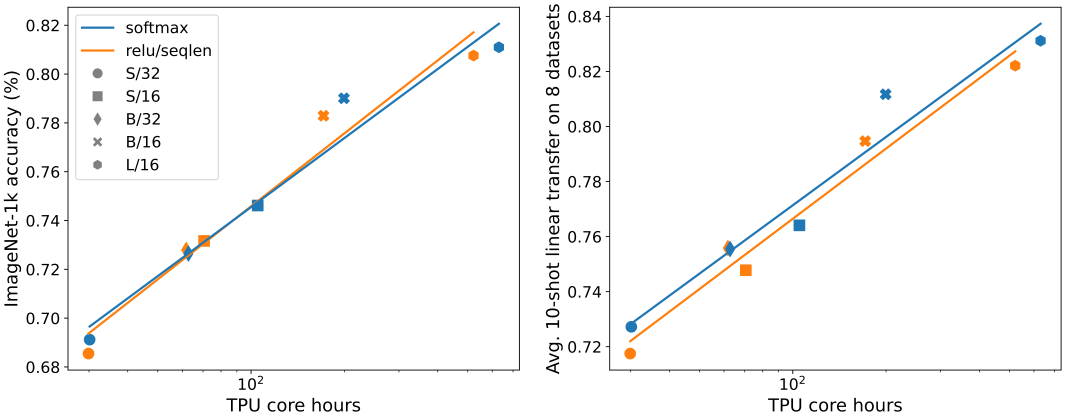

Figure 1: Replacing softmax with relu/seqlen approaches or matches the scaling performance of traditional attention for vision transformers [9] with qk-layernorm [7]. This figure displays results for small to large vision transformers trained on ImageNet-21k [8] for 30 epochs. We report ImageNet-1k accuracy for ImageNet-21k models by taking the top class among those that are in ImageNet-1k, without fine-tuning. Attention with ReLU can be parallelized over the sequence length dimension with less gather operations than softmax attention.

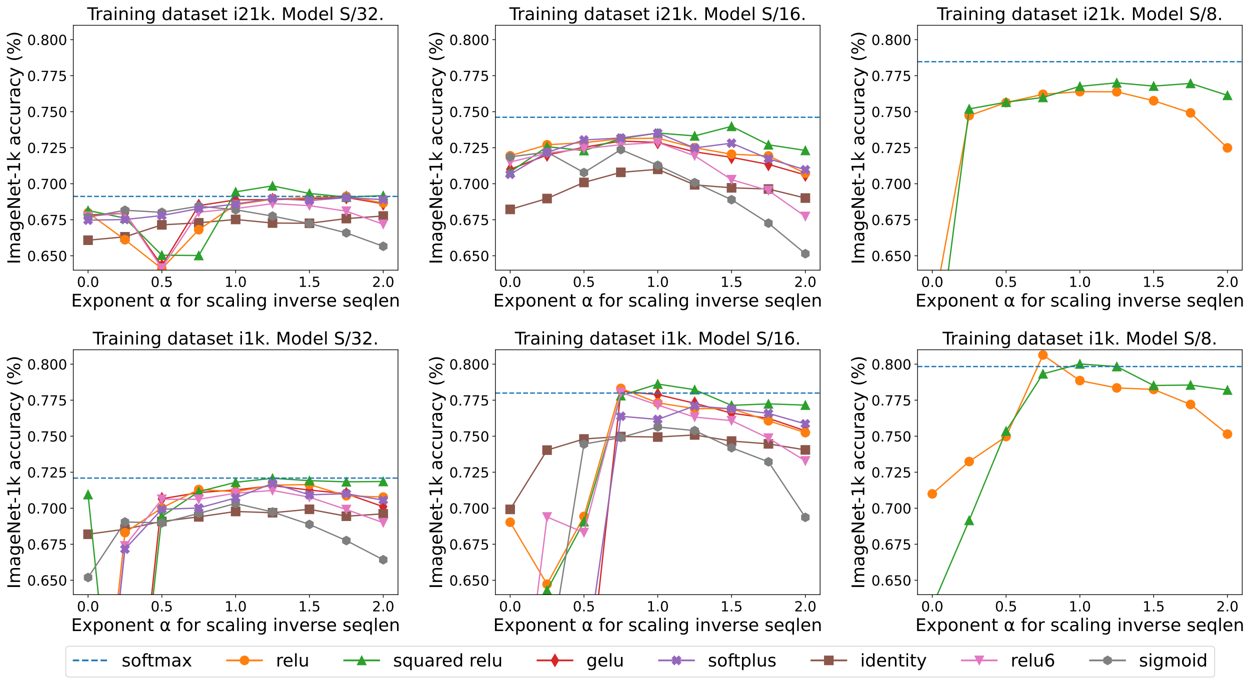

Figure 2: Replacing softmax with $L^{-\alpha}h$ where $h \in \{\mathsf{relu}, \mathsf{relu}^2, \mathsf{gelu}, \mathsf{softplus}, \mathsf{identity}, \mathsf{relu6}, \mathsf{sigmoid}\}$ and L is sequence length. We typically observe the best results when α is close to 1. There is no clear best non-linearity at $\alpha \approx 1$, so we use ReLU in our main experiment for its speed.

the attention weights sum to one. This retains the downside of requiring a gather.

Moreover, there is extensive literature which removes activation functions altogether so that attention is linear [14, 20, 16], which is useful for long sequence lengths.[1] In our experiments, removing the activation entirely reduced accuracy.

3 Method

Attention. Attention transforms d-dimensional queries, keys, and values $\{q_i, k_i, v_i\}_{i=1}^{L}$ with a two step procedure. First, attention weights α_{ij} are produced via

$$\alpha_{ij} = \phi\left(\frac{1}{\sqrt{d}}\left[q_i^\top k_1, ..., q_i^\top k_L\right]\right)_j, \tag{1}$$

where ϕ is typically $\mathsf{softmax}$. Next, the attention weights are used to compute outputs $o_i = \sum_{j=1}^{L} \alpha_{ij} v_j$. This report explores point-wise alternatives to ϕ.

ReLU-attention. We observe that $\phi = L^{-1}\mathsf{relu}$ is a promising alternative to $\phi = \mathsf{softmax}$ in Equation 1. We refer to attention with $\phi = L^{-1}\mathsf{relu}$ as ReLU-attention.

Scaled point-wise attention. More generally, our experiments will explore $\phi = L^{-\alpha}h$ for $\alpha \in [0, 1]$ and $h \in \{\mathsf{relu}, \mathsf{relu}^2, \mathsf{gelu}, \mathsf{softplus}, \mathsf{identity}, \mathsf{relu6}, \mathsf{sigmoid}\}$ [5, 11].

Sequence length scaling. We observe that scaling by a term involving sequence length L is beneficial for high accuracy. This scaling is absent from prior work which removes softmax [13, 16]. While the central justification for sequence length scaling is empirical, we provide brief analytical motivation.

Transformers are currently designed with softmax attention for which $\sum_{j=1}^{L} \alpha_{ij} = 1$. This implies that $\mathbb{E}_j[\alpha_{ij}] = L^{-1}$. While it is unlikely that this is a necessary condition, $\phi = L^{-1}\mathsf{relu}$ does ensure that $\mathbb{E}_j[\alpha_{ij}]$ is $O(L^{-1})$ at initialization. Preserving this condition may alleviate the need to change other hyperparameters when replacing softmax.

At initialization the elements of q and k are $O(1)$

[1] Concretely, with linear attention, the order of matrix multiplies can be switched from $(qk^\top)v$ to $q(k^\top v)$ which changes the compute required from $O(dL^2)$ to $O(d^2L)$ where $q, k, v \in \mathbb{R}^{L \times d}$ are the queries, keys, and values and L is sequence length.

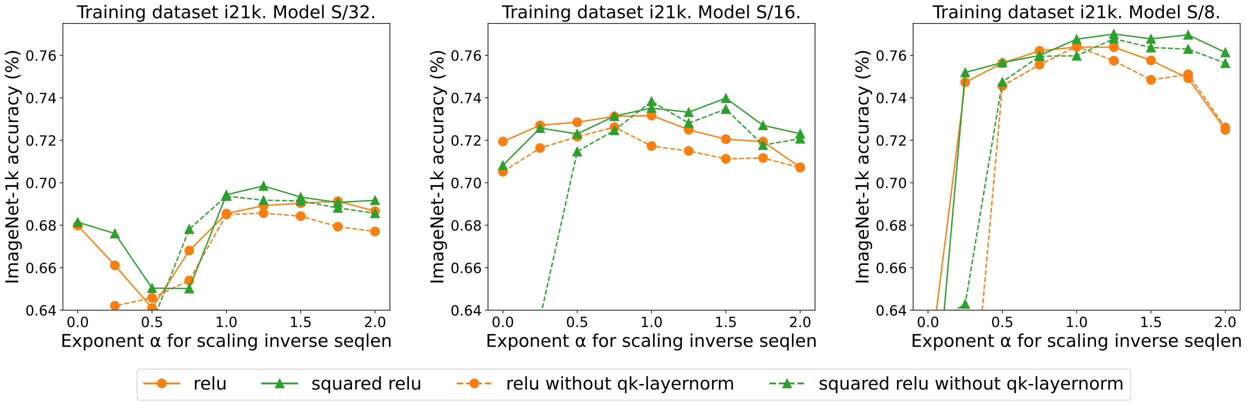

Figure 3: The effect of removing qk-layernorm [7] on attention with ReLU and squared ReLU scaled by $L^{-\alpha}$ where L is sequence length. Results are shown for the S/32, S/16, and S/8 vision transformer models [9, 3] trained on ImageNet-21k.

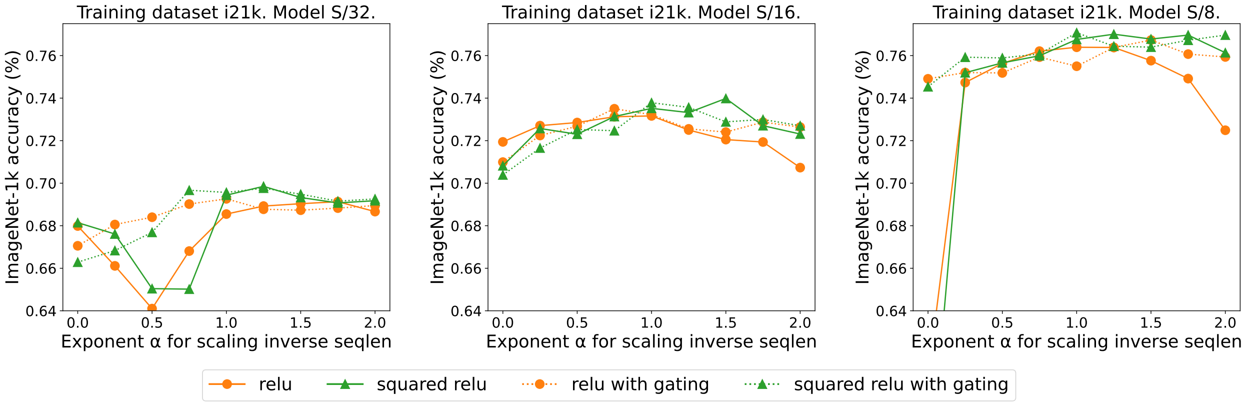

Figure 4: The effect of using a gated attention unit [13] on attention with ReLU and squared ReLU scaled by $L^{-\alpha}$ where L is sequence length. Results are shown for the S/32, S/16, and S/8 vision transformer models [9, 3] trained on ImageNet-21k.

and so $\frac{\langle q_i, k_j \rangle}{\sqrt{d}}$ will also be $O(1)$. Activation functions such as ReLU preserve $O(1)$,[2] and so a factor L^{-1} is necessary for $\mathbb{E}_j[\alpha_{ij}]$ to be $O(L^{-1})$.

4 Experiments

Experimental setup. Our experiments use ImageNet-21k and ImageNet-1k [8] training configurations from the BigVision codebase [3] without modifying hyperparameters.[3] In our experiments on ImageNet-21k we train for 30 epochs, and in our experiments on ImageNet-1k we train for 300 epochs. As a result, both training runs use a roughly similar number of steps of around 9e5. We use ViTs with qk-layernorm [7] as this was previously observed to be necessary to prevent instability when scaling model size. However, we ablate that this is not an important component at the scales we test. We use i21k and i1k to mean ImageNet-21k and ImageNet-1k, respectively, and report ImageNet-1k accuracy for ImageNet-21k models by taking the top class among those that are in ImageNet-1k, without fine-tuning. When evaluating transfer performance on downstream tasks we use a 10-shot linear probe averaged over three seeds. The downstream tasks are Caltech Birds [25], Caltech-101 [10], Stanford Cars [17], CIFAR-100 [18], DTD [4],

[2] With the exception of squared ReLU.

[3] For ImageNet1k we use the base config https://github.com/google-research/big_vision/blob/main/big_vision/configs/vit_i1k.py. For ImageNet21k we use the base config https://github.com/google-research/big_vision/blob/main/big_vision/configs/vit_i21k.py.

ColHsit [15], Pets [21], and UC Merced [26].

Main experiment. Figure 1 illustrates that ReLU-attention matches the scaling trends for softmax attention for ImageNet-21k training. On the x-axis we display the total core hours required for the experiment. As an advantage, ReLU-attention enables parallelization over the sequence length dimension with fewer gather operations than softmax attention.

Effect of sequence length scaling. Figure 2 examines the effect of sequence length scaling for various point-wise alternatives to softmax. Concretely, we replace softmax with $L^{-\alpha}h$ for $\alpha \in [0, 1]$ and $h \in \{\mathsf{relu}, \mathsf{relu}^2, \mathsf{gelu}, \mathsf{softplus}, \mathsf{identity}\}$. On the x-axis we display α. The y-axis displays accuracy for the S/32, S/16, and S/8 vision transformer models [9, 3]. The best results are typically achieved when α is close to 1. Since there is not clear best non-linearity, we use ReLU in our main experiment as it is faster.

Effect of qk-layernorm. Our main experiments use qk-layernorm [7] in which queries and keys are passed through LayerNorm [1] before computing attention weights. We use qk-layernorm by default as it was found to be necessary to prevent instability when scaling up model size [7]. Figure 3 shows the effect of removing qk-layernorm. The results indicate that qk-layernorm does not have a large effect for these models, but this may change at scale.

Effect of adding a gate. Previous work removing softmax adds a gated unit and does not scale by sequence length [13]. Concretely, in the gated attention unit [13] an extra projection produces output which is combined through elementwise-multiplication before the out projection. In Figure 4 we investigate whether the presence of a gate removes the need for sequence length scaling. Overall we observe that the best accuracy is still achieved with sequence length scaling, with or without the gate. Note that gating increases the core hours required for the experiment by roughly 9.3% for the S/8 model with ReLU.

5 Conclusion

This report leaves many open questions. In particular, we are unsure why the factor L^{-1} improves performance or if this term could be learned. Moreover, it is likely that there is a better activation function that we do not explore.

Acknowledgements

We thank Lucas Beyer, Mostafa Dehghani, and David Fleet for their helpful comments and suggestions.

We thank the members of the Google DeepMind PAGI team for their support of this effort, Jascha Sohl-dickstein, Noah Fiedel, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Avi Singh, Azade Nova, Ben Adlam, Bernd Bohnet, Daniel Freeman, Gamaleldin Elsayed, Gaurav Mishra, Hanie Sedghi, Isabelle Simpson, Izzeddin Gur, JD Co-Reyes, James Harrison, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kelvin Xu, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Max Bileschi, Merrie Morris, Roman Novak, Rosanne Liu, Sharad Vikram, Tris Warkentin, Yundi Qian.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022. URL https://arxiv.org/abs/2205.01580.

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. https://arxiv.org/abs/1311.3618.

[5] George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8609–8613. IEEE, 2013.

[6] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[7] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. *arXiv preprint arXiv:2302.05442*, 2023.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009. https://ieeexplore.ieee.org/document/5206848.

[9] 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. In *International Conference on Learning Representations (ICLR)*, 2021. https://arxiv.org/abs/2010.11929.

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[11] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[12] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks. In *International Conference on Machine Learning*, pages 4376–4386. PMLR, 2020.

[13] Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In *International Conference on Machine Learning*, pages 9099–9117. PMLR, 2022.

[14] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

[15] Jakob Nikolas Kather, Frank Gerrit Zöllner, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Cleo-Aron Weis. Collection of textures in colorectal cancer histology. *Zenodo https://doi. org/10*, 5281, 2016.

[16] Soroush Abbasi Koohpayegani and Hamed Pirsiavash. Sima: Simple softmax-free attention for vision transformers. *arXiv preprint arXiv:2206.08898*, 2022.

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International Conference on Computer Vision (ICCV) Workshops*, 2013. https://ieeexplore.ieee.org/document/6755945.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[19] Zhiyuan Li, Srinadh Bhojanapalli, Manzil Zaheer, Sashank Reddi, and Sanjiv Kumar. Robust training of neural networks using scale invariant architectures. In *International Conference on Machine Learning*, pages 12656–12684. PMLR, 2022.

[20] Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. Soft: Softmax-free transformer with linear complexity. *Advances in Neural Information Processing Systems*, 34:21297–21309, 2021.

[21] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[22] Markus N Rabe and Charles Staats. Self-attention does not need $o(n^2)$ memory. *arXiv preprint arXiv:2112.05682*, 2021.

[23] Kai Shen, Junliang Guo, Xu Tan, Siliang Tang, Rui Wang, and Jiang Bian. A study on relu and softmax in transformer. *arXiv preprint arXiv:2302.06461*, 2023.

[24] 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*, 30, 2017.

[25] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[26] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010.