SkipViT: Speeding Up Vision Transformers with a Token-Level Skip Connection

Foozhan Ataiefard, Walid Ahmed, Habib Hajimolahoseini, Saina Asani, Farnoosh Javadi, Mohammad Hassanpour, Austin Wen, Omar Mohamed Awad, Kangling Liu, Yang Li

Ascend Team, Huawei Toronto Research Centre

Abstract

Vision transformers are known to be more computationally and data-intensive than CNN models. These transformer models such as ViT [3], require all the input image tokens to learn the relationship among them. These tokens are overlooked by the multi-head self-attention (MHSA), resulting in many redundant and unnecessary computations in MHSA and the feed-forward network (FFN). In this work, we propose a method to optimize the amount of unnecessary interactions between unimportant tokens by separating and sending them through a different low-cost computational path. Our method does not add any parameters to the ViT model and aims to find the best trade-off between training throughput and achieving a 0% loss in the Top-1 accuracy of the final model. Our experimental results on training ViT-small from scratch show that SkipViT is capable of effectively dropping 55% of the tokens while gaining 13.23% training throughput and maintaining classification accuracy at the level of the baseline model on Huawei Ascend910A.

Background

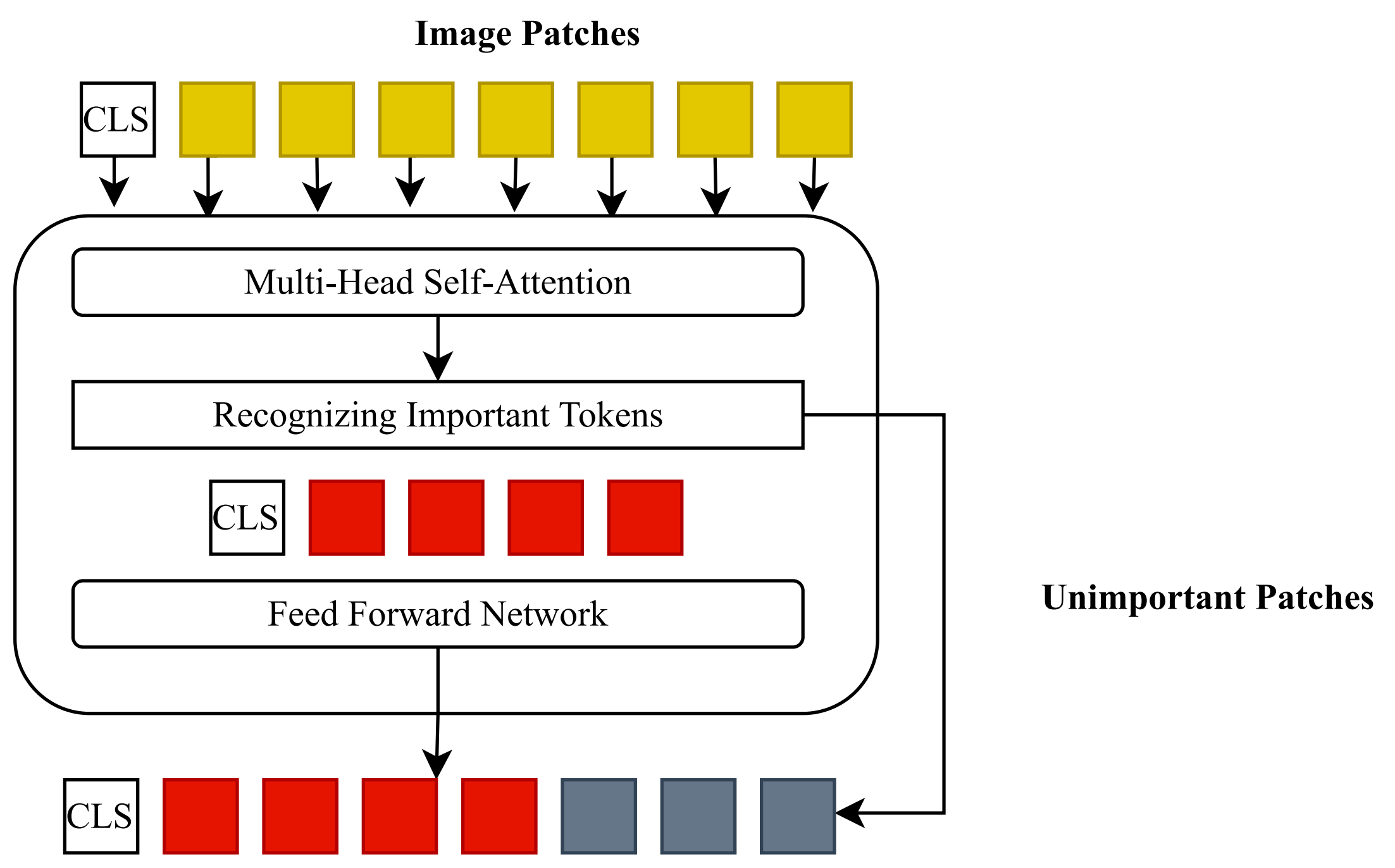
Multi-Head Attention (MHA). A crucial component in Transformer models [1], is designed to capture diverse aspects of the input data. The token inputs x to the attention layer are transformed into three distinct matrices: queries Q, keys K, and values V. These transformations are achieved through a linear transformation.

The attention mechanism in each head is computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \qquad (1)$$

where d represents the dimensionality of the key (and query) vectors.

Here, QK^T is the dot-product of queries and keys, and $\sqrt{d}$ is the scaling factor to avoid large values in the dot-product attention. In this paper we refer to the resulting matrix of $Softmax(QK^t/\sqrt{d})$ as the attention scores.



Figure

Proposed Method

Identifying Important Patches. In ViT models, we can employ the attention scores corresponding to `[CLS]` token to detect important patches of an image [2]. The attention scores consist of a $(n + 1) \times (n + 1)$ matrix where n is the number of input tokens to the attention unit. Based on the attention Eq. 1, we can say that each row i in the attention score matrix are coefficients by which other tokens will attend in forming the new i token at the attention unit output.

Skip Connection For Tokens. By dropping 45% of the tokens from the 6th transformer layer of ViT-small and adding a single fused token, which incorporates a weighted average of the removed tokens, our model was unable to maintain baseline accuracy while gaining throughput, as shown in Table 1. We propose the use of a skip connection for the tokens that would otherwise be discarded. This approach selectively excludes these tokens from contributing to certain transformer layers within the model, while still incorporating them in the final layers. Returning the dropped tokens to their original position among other tokens reduces the impact of token dropping on final classification accuracy of the model.

Experiments

We performed all of our experiments using ViT[4] as our baseline architecture. We experimented with two strategies for dropping the tokens. A single and a two stage token dropping (i.e., drop in one or two layers) strategy to find the best trade-off between training performance and final accuracy of the ViT model between these methods. A summary of our experimental results are presented in Table 1.

In both of the dropping methods we were able to see a relative speedup with limited to no loss in the validation accuracy. In single layer token dropping method, our best method with dropping 55% of the tokens at layer 6 with skip connection to layer 11 reaches 0.01% accuracy drop while gaining 13.23% throughput. Using two stage token dropping approach and drop ratio of 30% for layers 4 and 7 with skip connection to layer 11, our fastest achieved 16.09% more FPS and reaching 69.4% classification accuracy which outperforms the token fusion technique.

Dropping	Throughput	Top-1(%)
ViT-small	4,503	70.17
6(45%)+fused token	4,963(+10.23%)	68.39(-1.78)
4,7(30%,30%)+10	5,092(+13.1%)	69.73(-0.44)
4,7(30%,30%)+11	**5,227(+16.09%)**	69.4(-0.77)
6,8(35%,35%)+10	4,711(+4.62%)	70.53(+0.36)
6,8(35%,35%)+11	4,838(+7.45%)	70.41(+0.24)
6(45%)+11	4,944(+9.8%)	**70.64(+0.47)**
6(50%)+11	5,021(+11.51%)	70.27(+0.1)
6(55%)+11	5,098(+13.23%)	70.16(-0.01)

Finding The Optimal Skip Connection To prevent from any degradation in Top-1 accuracy of the ViT model we reuse the dropped tokens in the future layers. Based on our findings delaying the skip connection by even 1 block can cause a substantial decrease in the accuracy metric. Dropping 30% of the tokens at layers 4 and 7 and returning them to the sequence at 10th layer compared to returning at 11th layer, achieves 2.99% higher FPS while loosing 0.33% accuracy.

Effect of Warm-up On Patch Detection Quality We also experimented with a warm-up period for this approach. Asimilar dropping ratio (30%) in the same layers (4 and 7), ViT reaches 2.55% higher accuracy when first 15 epochs are used as warm-up period before token dropping is applied. We saw that the warm-up epochs are a essential part of our token dropping strategy which helps the model to select a more informative set of tokens to keep.

Conclusion

In this paper, we propose SkipViT, an intuitive and stable framework to effectively reduce the amount of computation required to train ViT-based models. SkipViT takes advantage of the attention scores of the `[CLS]` token to differentiate the computation graph between important from less informative tokens. Furthermore, Our proposed framework achieves a significant speedup with no loss in the accuracy of the model by adding a skip connection from the dropping block to a future transformer block in ViT. This method shows promising results on the current setup, However it is limited by the size of the model and dataset and should be extended to the larger versions of ViT.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need

[2] Liang, Youwei, G. E. Chongjian, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. 2021. "EViT: Expediting Vision Transformers via Token Reorganizations." In International Conference on Learning Representations.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.