An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale

Summary written by Nicholas Hinke
April 19, 2022

Summary: As demonstrated by a variety of authors throughout recent years, the importance and prevalence of self-attention-based architectures has been growing rapidly, especially within the Natural Language Processing (NLP) community. In fact, Transformers in particular have arguably become the most dominant area of research within the field. While there exist many reasons for this sudden rise in popularity, chief among them are the computational efficiency and scalability of Transformers, which allow for the construction of incredibly complex models. Despite their popularity in the NLP community, however, Transformers had gained little traction among computer vision researchers. In order to remedy this potentially missed opportunity, the authors of this paper developed what is known as the "Vision Transformer" (ViT) [4]. Heavily inspired by other recent works regarding self-attention mechanisms [2, 8, 10], model pretraining [1, 3], and successful image recognition models [5, 6, 11], the authors sought to develop a Transformer capable of outperforming other state-of-the-art image recognition methods. Indeed, when supplying the model with sufficient training data (many millions of samples) and a suitable pretraining routine, the resulting Transformer model was able to compete with many other state-of-the-art methods on a variety of image recognition benchmarks [4].

Approach: In contrast to previous works that attempted to combine traditional convolutional approaches with self-attention [2, 7, 9, 10], the authors of this paper sought to emulate the original Transformer design [8] as closely as possible. In that vein, the Transformer encoder itself is structured in a similar manner to that of the original architecture, where the one-dimensional Transformer inputs are embeddings of fixed-size image patches. In order to construct these embeddings, a two-dimensional image is first reshaped into a sequence of flattened patches. These patches are then linearly embedded via a trainable linear projection, before trainable one-dimensional position embeddings are added in order to retain each patch's positional information. These embedded patches are then fed into the Transformer, where they pass through alternating layers of multiheaded self-attention (MSA) and multi-layer perceptron (MLP) blocks. It should also be noted that layernorm is applied before each block, and residual connections are utilized after each block. Finally, the Transformer outputs are passed through a simple MLP for classification purposes [4].

In order to fully take advantage of the Transformer architecture–primarily due to the typically very large size and complexity–it is necessary for the ViT to first be pretrained on large datasets. Following adequate pretraining, transfer learning can be employed to fine-tune the network on the task at hand. During their experiments, the authors considered ViT models of size varying from 86 million parameters to 632 million parameters, which were pretrained on 2012 ImageNet (1.3 million images), ImageNet-21k (14 million images), or JFT (303 million images). The resulting Transformers were then fine-tuned on a variety of image recognition benchmarks and compared to other state-of-the-art ResNet models and hybrids. In fact, the largest ViT model pretrained on JFT outperformed all of its competitors in almost every benchmark category (6/7) [4, 5].

Strengths: Most notably, the resulting ViTs as constructed by the authors were able to outperform many other state-of-the-art image recognition methods, thus demonstrating the true potential of Transformers within the field of computer vision. Furthermore, the authors were able to provide a more intuitive representation of the potential benefits of attention mechanisms through easily understood visualizations [4]. Additionally, due to the decision of the authors to follow as closely as possible to the original Transformer design [8], it is quite straightforward to instead implement a variety of other NLP Transformer architectures within their ViT framework. Finally, the authors likely influenced several future related advances in the field, as they made their trained models publicly available online [4].

Weaknesses: While the authors cite several examples of the successes achieved by Transformers of a much greater size and complexity, they made no attempt to construct such a ViT. Considering the success of their largest ViT, this seems like an oversight that may have cost them significant performance improvements. Additionally, while mentioned as potential avenues of future research, the authors fail to provide any recommendations or discussion regarding adaptation of ViTs to other computer vision tasks such as segmentation or detection [4].

Reflections: Briefly mentioned within the paper, the authors experimentally demonstrated the potential success of self-supervised model pretraining. In that vein, further research could prove immensely beneficial in both the model performance and simplicity of training. Furthermore, additional study involving the implementation of other related NLP Transformer architectures within the proposed ViT framework may provide further insights and advancements within the field of computer vision [4].

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

[5] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2020.

[6] D. K. Mahajan, R. B. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *ArXiv*, abs/1805.00932, 2018.

[7] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, 2019.

[8] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[9] H. Wang, Y. Zhu, B. Green, H. Adam, A. L. Yuille, and L.-C. Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, 2020.

[10] X. Wang, R. B. Girshick, A. K. Gupta, and K. He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[11] Q. Xie, E. H. Hovy, M.-T. Luong, and Q. V. Le. Self-training with noisy student improves imagenet classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020.