Trimble / Bilberry : AI Engineer technical exercise Part 2 : Paper Review

Chappe Aslan
April 2023

The research article "*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*" is dealing with the use of transformers for image recognition, a field that has previously been dominated by convolutional neural networks (CNN). The authors explain that transformers can also be successful in picture identification tasks.

The Vision Transformer (ViT), a brand-new architecture that the authors suggest can be used to interpret pictures, combines transformers with convolutional layers. The picture is divided into patches by the ViT, and each patch is afterwards processed by a transformer. The final output is created by running the created representation through a number of convolutional layers.

According to the authors, ViT networks have better performances than CNN on well known image datasets such as ImageNet and CIFAR-100. This is at the moment the model that achieves state-of-the-art performance on the ImageNet. What is remarkable about ViT is that it doesn't need the images to be the same size to process them, as the opposite of CNN that required some steps of resizing and preprocessing. It leads to a model not having to make sacrifices over performance or speed.

However, the ViT is computationally and memory-intensive, which may limit its applicability in certain situations. The article proposes several techniques for reducing the computational cost of the ViT, including distillation and efficient attention mechanisms.

Overall, the authors argue that the ViT represents a significant step forward in the application of transformers to image recognition tasks. The success of the ViT suggests that transformers hold great promise for image recognition at scale and may open new ways for research in this area.

Source

Article : Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv, abs/2010.11929*.

https://arxiv.org/abs/2010.11929