A COMPREHENSIVE REVIEW ON TRANSFORMERS NEURAL NETWORKS

Dr. B. Gnana Priya

Assistant Professor Department of Computer Science and Engineering,
Faculty of Engineering and Technology
Annamalai University

ABSTRACT

A transformer is a network of nodes which learns some task by training on an existing set of data. Initially transformers were designed to perform natural language processing. In recent times transformers were found to give better results for many of the artificial intelligence and computer vision problems. The neural networks and deep learning approaches first learn from the local patches of input data and then will try to build up to the whole system. The transformer by contrast runs processes so that every element in the input data connects or pays attention to every other element and this property is known as self-attention. This means that as soon as it starts training, the transformer can see traces of the entire data set. Transformers are proving surprisingly versatile. In some vision tasks, like image classification, neural nets that use transformers have become faster and more accurate than those that don't. Emerging work in other AI areas like processing multiple kinds of input at once or planning tasks suggests transformers can handle even more.

1. INTRODUCTION

The transformer first appeared in 2017 in a paper that cryptically declared that "Attention Is All You Need. While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. It is proved that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks like ImageNet, CIFAR-100, VTAB, etc., Vision Transformer attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

2. RELATED WORK

Transformers were proposed by Vaswani et al. (2017) for machine translation, and have since become the state of the art method in many NLP tasks. Large Transformer-based models are often pre-trained on large corpora and then fine-tuned for the task at hand: BERT (Devlin et al., 2019) uses a denoising self-supervised pre-training task, while the GPT line of work uses language modeling as its pre-training task (Radford et al., 2018;). Naive application of self-attention to images would require that each pixel attends to every other pixel. With quadratic cost in the number of pixels, this does not scale to realistic input sizes. Thus, to apply Transformers in the context of image processing, several approximations have been tried in the past. Parmar et al. (2018) applied the self-attention only in local neighborhoods for each query pixel instead of globally. Such local multi-head dot-product self attention blocks can completely replace convolutions (Hu et al., 2019;). In a different line of work, Sparse Transformers (Child et al., 2019) employ scalable approximations to global selfattention in order to be applicable to images. An alternative way to scale attention is to apply it in blocks of varying sizes (Weissenborn et al., 2019), in the extreme case only along individual axes (Ho et al., 2019; Wang et al., 2020a). Many of these specialized attention architectures demonstrate promising results on computer vision tasks, but require complex engineering to be implemented efficiently on hardware accelerators. Most related to ours is the model of which extracts patches of size 2 × 2 from the input image and applies full self-attention on top. This model is very similar to ViT, but our work goes further to demonstrate that large scale pre-training makes vanilla transformers competitive with (or even better than) state-of-the-art CNNs. Moreover, Cordonnier et al. (2020) use a small patch size of 2 × 2 pixels, which makes the model applicable only to small-resolution images, while we handle medium-resolution images as well. There has also been a lot of interest in combining convolutional neural networks (CNNs) with forms of self-attention, e.g. by augmenting feature maps for image classification (Bello et al., 2019) or by further processing the output of a CNN using self-attention, e.g. for object detection (Carion et al., 2020), video processing (Wang et al., 2018; Sun et al., 2019), image classification (Wu et al., 2020), unsupervised object discovery, or unified text-vision tasks (Chen et al., 2020c;). Another recent related model is image GPT (iGPT) (Chen et al., 2020a), which applies Transformers to image pixels after reducing image resolution and color space. The model is trained in an unsupervised fashion as a generative model, and the resulting representation can then be fine-tuned or probed linearly for classification performance, achieving a maximal accuracy of 72% on ImageNet. Our work adds to the increasing collection of papers that explore image recognition at larger scales than the standard ImageNet dataset. The use of additional data sources allows to achieve state-ofthe-art results on standard benchmarks (Mahajan et al., 2018; Touvron et al., 2019; Xie et al., 2020). Moreover, Sun et al. (2017) study how CNN performance scales with dataset size, and Kolesnikov et al. (2020); Djolonga et al. (2020) perform an empirical exploration of CNN transfer learning from large scale datasets such as ImageNet-21k and JFT-300M. We focus on these two latter datasets as well, but train Transformers instead of ResNet-based models used in prior works.

3. TAXONOMY OF TRANSFORMERS

A wide variety of models have been proposed so far based on the vanilla Transformer from three perspectives: types of architecture modification (lightweight variants, cross-block connectivity, Adaptive Computation Time, recurrence & hierarchy, alternative architectures), pre-training methods, and applications(Text , Vision, Audio, Multimodal). There exists a variety of Transformer variants based on several characteristics. Categorization at module level based on i) attention (Sparse, Linearized, Prototype, Low Rank, Multihead, Prior Attention), ii) position encoding (absolute, relative, implicit), iii) layer normalization (placement, substitution, normalization tree) and iv) activation function exists. Architecture level categorization like lightweight (Lite Transformer), connectivity (Realformer, Feedback Transformer) and transformers based on divide and conquer (Transformer XL, Compressive Transformer) are used in various scenarios. Based on application many categories like BERT, transformer-XL for Natural Language Processing, Image transformer, ViT for computer vision, Speech transformer, Music transformer for audio and VisualBERT, VideoBERT for multimodal are available.

The vanilla Transformer is a sequence-to-sequence model and consists of an encoder and a decoder, each of which is a stack of L identical blocks. Each encoder block is mainly composed of a multi-head self-attention module and a position-wise feed-forward network (FFN). For building a deeper model, a residual connection is employed around each module, followed by Layer Normalization module. Compared to the encoder blocks, decoder blocks additionally insert cross-attention modules between the multi-head self-attention modules and the position-wise FFNs. Furthermore, the self-attention modules in the decoder are adapted to prevent each position from attending to subsequent positions.

4. ATTENTION

Transformer adopts attention mechanism with Query-Key-Value (QKV) model. Given the packed matrix representations of queries $\mathrm{Q} \in \mathrm{R}^{N \times}$, keys $\mathrm{K} \in \mathrm{R}^{M \times Dk}$, and values $\mathrm{V} \in \mathrm{R}^{M \times Dv}$, the scaled dot-product attention used by Transformer is given by

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\ (\mathrm{QK}^{\top} / \sqrt{}\ D_k)\ \ \mathrm{V} = \mathrm{AV} \qquad (1)$$

where N and M denote the lengths of queries and keys (or values); Dk and Dv denote the dimensions of keys (or queries) and values; $\mathrm{A} = \mathrm{softmax}\ (\mathrm{QK}^{\top} / \sqrt{}\ D_k)$ is often called attention matrix; softmax is applied in a row-wise manner. The dot-products of queries and keys are divided by $\sqrt{}\ D_k$ to alleviate gradient vanishing problem of the softmax function. Instead of simply applying a single attention function, Transformer uses multi-head attention, where the Dm-dimensional original queries, keys and values are projected into , Dk and Dv dimensions, respectively, with H different sets of learned projections. For each of the projected queries, keys and values, and output is computed with attention according to Equation (1). The model then concatenates all the outputs and projects them back to a Dm-dimensional representation.

$$\mathrm{MultiHeadAttn}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Concat}(\mathrm{head}_1, \cdot \cdot \cdot , \mathrm{head}_H)\mathrm{W}^{O} \qquad (2)$$

where $\mathrm{head}_i = \mathrm{Attention}(\mathrm{QW}^{Q}{}_{i}\ , \mathrm{KW}^{K}{}_{i}\ , \mathrm{VW}^{V}{}_{i}\ )$.

The advancements in attention mechanism can be divided into several categories based on: (1) Sparse Attention. This line of work introduces sparsity bias into the attention mechanism, leading to reduced complexity. (2) Linearized Attention. This line of work disentangles the attention matrix with kernel feature maps. The attention is then computed in reversed order to achieve linear complexity. (3) Prototype and Memory Compression. This class of methods reduces the number of queries or key-value memory pairs to reduce the size of the attention matrix. (4) Low-rank Self-Attention. This line of work capture the low-rank property of self-attention. (5) Attention with Prior. The line of research explores supplementing or substituting standard attention with prior attention distributions. (6) Improved Multi-Head Mechanism. The line of studies explores different alternative multi-head mechanisms.

5. CONCLUSION

The architecture of Transformer has been demonstrated to be capable of supporting large-scale training datasets with enough parameters. Many works show that Transformer has a larger capacity than CNNs and RNNs and hence has the ability to handle a huge amount of training data. Most of the existing works improve Transformer from different perspectives, such as efficiency, generalization, and applications. The improvements include incorporating structural prior, designing lightweight architecture, pre-training, and so on. Although X-formers have proven their power for various tasks, challenges still exist.

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. In NIPS, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2019.

[3] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. Technical Report, 2018.

[4] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In ICML, 2018.

[5] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In CVPR, 2018.

[6] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. arXiv, 2019.

[7] Dirk Weissenborn, Oscar Tackstr ¨ om, and Jakob Uszkoreit. Scaling autoregressive video models. In ¨ ICLR, 2019.

[8] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. arXiv, 2019.

[9] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In ECCV, 2020a

[10] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between selfattention and convolutional layers. In ICLR, 2020

[11] Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens. Attention augmented convolutional networks. In ICCV, 2019.

[12] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In ECCV, 2020c.

[14] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In ECCV, 2018.

[15] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Herve Jegou. Fixing the train-test resolution discrepancy. In NeurIPS. 2019

[16] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. In ECCV, 2020.

[17] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, Sylvan Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks. arXiv, 2020.