5 | 13 June 23

Topic: Escaping the big data paradigm with compact tranformers

Abstract:

With the rise of transformers as the standard for language processing and their advancements in computer vision, there has been a corresponding growth in parameter size and amount of training data. Many have come to believe that because of this, transformers are not suitable for small sets of data. This trend leads to concerns such as: limited availability of data in certain scientific domains and the exclusion of those with limited resource from research in the field. In this paper, we aim to present an approach for small-scale learning by introducing Compact Tranformers We show for the first time that for with the right size, convulational tokenization, transformers can avoid overfitting and outperform state of the art CNNs on small datasets. Our models are flexible in terms of modle size, and can have as little as 0.28M parameters while achiving competitive results. Our best model can reach 98% accuracy when training from scratch on CIFAR-10 with only 3.7M parameters, which is a significant impro in data efficiency ouer precious Transformer based models being over 10x smaller than other transformers and is 15% the size of ResNet50 while achieving similar performance, CCT also outperforms many modern CNN based approaches and even some recent NAS based approaches. Additionally, we obtain a new SOTA result on flowers-102 with 99.76% top-1 accuracy and improve upon the existing baseline on Imagenet as well as NLP tasks. our simple and compact design for transformers makes them more feasible to study for those with limited computing resources dealing with small datasets while extending research efforts in data efficient transformers

<u>Conclusion</u>:

Transformers have commonly been percieved to be only applicable to large scale or medium scale training. while their scalability is undeniable, we have shown within this paper that with proper configuration, a transformer can be used in small data regimes success-fully as well and outperform convulational models of equivalent and even larger sizes. Our method is simple, flexible in size and the smallest of our variants can be easily loaded on even a minimal GPU or even a ~~GPU~~ CPU. while part of ...ch has been focused on large scale models and data sets, we focus on smaller scales in which there is still much research to be done in data efficiency. we show that CCT can outperform other transformer based models on small datasets while also having a significant reduction in computational costs & memory constraints. This work demonstrates that transformers dont require vast computational resources and can allow for their applications in even the most modest of settings. This type of resach is important to many scientistific domains where data is far more limited that the conventional machine learning datasets which are used in general research continuing research in this direction will help open research up to more people and domains, extending machine learning research.