CS230

Vision Transformers using Message Passing/ Vector Attention

Brendan R. Evers
Department of Computer Science
Stanford University
brev0793@stanford.edu

1 Introduction

Vision transformers have come to dominate the world of higher-performance computer vision. Transformers are attention networks, making use of a simple scalar value to pass information between elements. It has been noted that transformers can be thought of as graph attention networks, and that message passing is an existing technique to pass more rich information between nodes of a graph. We propose a vision transformer that makes use of message passing as an alternative to the scalar attention of traditional transformers. By passing more information between the nodes of our graph, we might hope to achieve higher performance than traditional methods.

2 Background

Transformers were originally introduced by Vaswani et al. in the 2017 paper Attention is All You Need[1]. Initially applied to the machine translation task and building upon previous models that used both recurrent neural networks and attention, transformers eschew the recurrence component, relying entirely on a mechanism called multi-headed self attention. Self attention works by breaking the embedding of words up between multiple heads. These heads are then passed through three separate weight matrices to compute query, key, and value matrices. The queries and keys are multiplied in order to create attention scalars, which scale the values before they are passed to the next layer of the network.

Transformers were later applied to the computer vision task in the paper An Image is Worth 16x16 Words by Dosovitskiy et al[2]. Rather than using sentences composed of words for inputs, this paper used images by breaking the images into several patches, flattening them, and dropping them into the transformer architecture in the same manner as Vaswani had previously done, replacing sentences with images and word-embeddings with patches. This architecture achieved state-of-the-art level performance on image classification benchmarks in 2020. While alterations have been made to the architecture in the form of things like Google's MaxVit[3], the basic concept of using transformers for the image classification task remains state of the art.

In a lecture at Cambridge, Petar Veličković points out that transformers can be thought of as graph attention networks[4], where the words (or in our case, patches) function as nodes and the attentions function as attentions. In this project, we would like to expand on this idea, and attempt to apply techniques used by graph neural networks to improve the performance of vision transformers.

It is expressed in the same lecture that message passing networks are an alternative to graph attention networks that pass more rich information between the nodes of the network than attention alone. This technique, first proposed in the paper Neural Message Passing for Quantum Chemistry by Gilmer et al.[5], passes vectors between the nodes of the graph network rather than scalars, and is in principle more expressive than attention passing. Additionally, Message passing neural networks have

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

been used for transformer-like self-attention for the purpose of document understanding in Message Passing Attention Networks for Document Understanding by Nikolentzos et al.[6] We would like to apply this same framework to the task of image classification. Other aggregation techniques over graphs have been used in the past as well, like the pointwise maximums used by Pointnet[7], which we would like to test as well.

3 Dataset

We had intended to train on the well-known and established ImageNet dataset, which consists of 14,197,122 labeled colour images. We selected this dataset because it has served as one of the most popular benchmarks for image classification tasks for years, and we would like to compare our performance to well-understood models. Due to compute and time constraints, we used a subset of ImageNet called Tiny-ImageNet with 200 classes and 100,000 training images [8]. Each class has 500 training examples and 100 test examples, and consists of a 64x64 color image along with a class label. Some example images from Tiny-Imagenet is shown below.

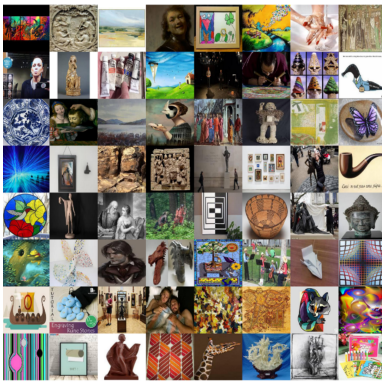

figure 1, Tiny-ImageNet sample

4 Method

We are comparing three different methods of attention: Conventional self-attention, and two vector attention techniques. These vector attention techniques are designed to treat the values of our patches like nodes in a graph, and pass richer information between them than scalar attention. We've decided to call the newly added dimension the "route" dimension. We're testing two methods of achieving this: With the first we're simply expanding the key and query outputs in one dimension while keeping the value matrix the same, and the taking the point-wise maximum of the resulting attention matrix along our new route dimension before multiplying it into value matrix, which is a technique inspired by the pointnet architecture [7]. Our message-passing network expands across both dimensions, giving both our attention and value matrices route dimensions. We then aggregate by taking the mean across our route dimension, matching our input to our output. We made use of an existing existing ViT PyTorch repository [9] to serve as the basis of our own repository.

5 Baseline and Baseline Results

The baseline model we hoped to imitate was the ViT-Base model provided by the Dosovitskiy paper[2]. We chose this as it was the smallest model the authors of the original paper provided, and we felt constrained on compute resources. As a result, our baseline model had 12 heads, 12 layers, a hidden dimension size of 768, and an MLP size of 3072. This gave us 86 million parameters. While exact performance for this model wasn't provided by the paper, we felt it likely that it would provide comparable performance to the 88.62 accuracy seen in the paper[3] In practice, we found that our model produced considerably inferior results, yielding an accuracy of only 30.11. We made use of the Adam optimizer, and swept our learning rate by powers of 10 from 1e-5 to 1e-2 when tuning our hyperparameters.

6 Message Attention and Message Attention Results

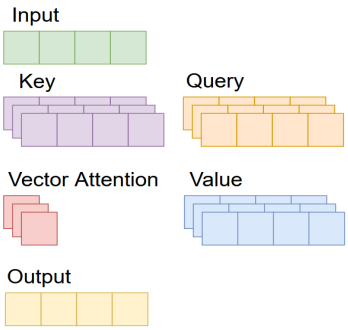

figure 2, Message Attention Architecture

Unlike conventional attention, which scales the value vector of a transformer, our Message Attention architecture produces a vector attention and a three-dimensional value tensor. This yields the same output shape, but combines multiple attentions and values along our newly-introduced route axis. In practice, we scale the output down by a factor of the route dimension in order to avoid growth of values between layers. We achieve these expanded vectors by expanding the dimension of our key, query, and value weight matrices in the route dimension, to produce the three-dimensional key, query, and value tensors shown above. When selecting parameters for model size we imitated the ViT-Base model entirely with the single exception of the route parameter. We varied the size of the route parameter between 4 and 16 and found the best performance at 4. Using this we achieved an accuracy of 28.01.

7 Max Attention

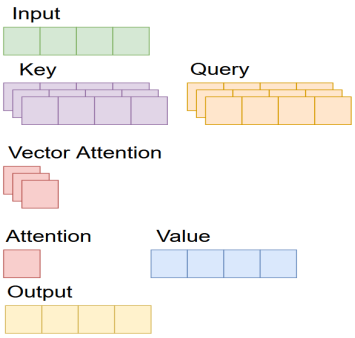

figure 3, Max Attention Architecture

Max attention is an additional method we tried, inspired by the PointNet architecture. Rather than using a full message as with Message Attention, we only expanded the query and key matrices. We then take a pointwise-maximum of our attention vector, and scale our value vector as normal, without any additional normalization. Once again, when selecting parameters for model size we imitated the ViT-Base model entirely with the single exception of the route parameter. We varied the size of the route parameter between 4 and 16 and found the best performance at 4. Using this we achieved an accuracy of 26.70.

Table of results

Model	Baseline	Message	Max
Accuracy	30.11	28.01	26.70

8 Analysis of Results

One of the most striking aspects of our results is the relatively low accuracy achieved by all three models. We believe that this is due to scaling laws. While the Base-ViT is provided by the Dosovitskiy paper, it has less than a third the number of parameters of the models for which results are actually presented. In addition, our constrained training set is only about 1/140th the size of the ImageNet dataset. Kaplan believe that scaling laws in both model size and training data are known to exist in transformers applied to the natural language processing task[10], and we believe they also apply to the vision task. This explains our strangely low accuracy.

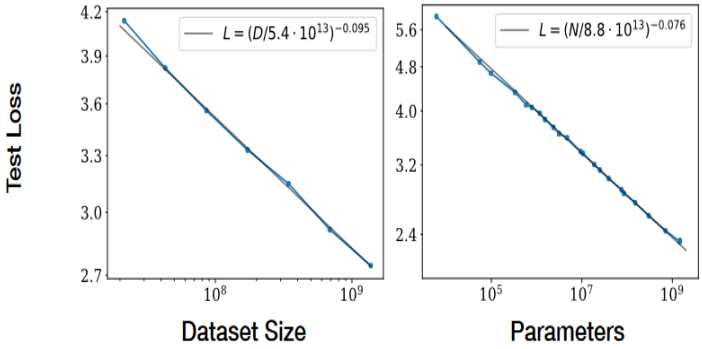

figure 4, taken from [10]

With regards to the very low variation in accuracy and indeed lower performance of our novel attention methods, it seems that our hypothesis that more expressive forms of attention could lead to superior results was incorrect. This is likely because, as Vaswani et al. suggest in 2017, attention is really all you need[1]. Attention already passed sufficient information between the nodes, and all our additional features did was serve to slow model convergence. Indeed, a significant challenge posed by the max architecture is that the gradients for all of the components not selected by the pointwise-maximum are zero, and it may indeed eventually perform as well as the baseline if given enough training time. Our message passing network seems to add additional complexity to the function the Transformer needs to learn to no additional benefit, and simply reduces our performance compared to baseline, otherwise functioning as a normal image transformer.

[1] Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin Attention Is All You Need 2017 arXiv:1706.03762

[2] Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and Xiaohua Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale 2020 arXiv:2010.11929

[3] Zhengzhong Tu and Hossein Talebi and Han Zhang and Feng Yang and Peyman Milanfar and Alan Bovik and Yinxiao Li MaxViT: Multi-Axis Vision Transformer 2022 arXiv:2204.01697v4

[4] Ellis Unit (June 14 2022)Petar Veličković Graph Neural Networks: Geometric, Structural and Algorithmic Perspectives Part 2 [Video]. YouTube. https://www.youtube.com/watch?v=pL5Nc8Axv5At=2418s

[5] Neural Message Passing for Quantum Chemistry J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl. Proceedings of the 34th International Conference on Machine Learning, Vol 70, pp. 1263–1272. PMLR. 2017.

[6] Giannis Nikolentzos and Antoine J.-P. Tixier and Michalis Vazirgiannis Message Passing Attention Networks for Document Understanding 2022 arXiv 1908.06267v2

[7] Charles R. Qi and Hao Su and Kaichun Mo and Leonidas J. Guibas PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation 2016, arXiv:1612.00593

[8] https://huggingface.co/datasets/Maysee/tiny-imagenet

[9] https://github.com/lucidrains/vit-pytorch

[10] Jared Kaplan and Sam McCandish and Tom Henighan and Tom Brown and Benjamin Chess and Reown Child and Scott Gray and Alec Radford and Jeffrey Wu and Dario Amodei Scaling Laws for Neural Language Models 2020 arXiv 2001.08361v1