**Vision Transformer Models of GT concepts**

Members – Aishwarya Solanki, Nikhil Sachdeva

**Abstract**

Vision Transformer (ViT) models have revolutionized computer vision by delivering exceptional performance across tasks like classification and object detection, often surpassing state-of-the-art CNNs. Unlike CNNs, which rely on localized receptive fields, ViTs leverage self-attention mechanisms to capture global context. In this study, we propose to explore whether the benefits of using transformer architectures offer any advantages in mathematical cognition, specifically in the domains of Geometry and Topology (GT) concepts.

Building on the foundational work of *Dehaene et al.* on the Amazonian Mundurukú tribe and *Vijay et al.’s* subsequent experiments, this study aims to replicate and extend their findings using newer state-of-the-art ViT architectures. The project incorporates a “checkpointing” approach, enabling incremental model training and saving snapshots at regular intervals.

The data used in this study will be the images from the study published by *Dehaene et al.* containing the odd-one-out tasks with six images each in which five of them demonstrate a particular GT concept whereas the sixth image does not. These tasks contain items for 43 individual GT concepts broadly grouped into seven superclasses - Topology, Euclidean Geometry, Geometrical Figures, Symmetrical Figures, Chiral Figures, Metric Properties, and Geometrical Transformations.

Performance metrics such as accuracy and F1 scores will be analysed across checkpoints to understand how the model learns GT concepts, and identify whether we can pinpoint any developmental characteristics for these concepts as the model trains.

Potential ViT model architectures to be explored - [BEiT, DINO, MAX-ViT, ConvNext] [4] [5] [6] [7]

**References**

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
2. Stanislas Dehaene *et al.* Core Knowledge of Geometry in an Amazonian Indigene Group.Science311,381-384(2006). DOI:[10.1126/science.1121739](https://doi.org/10.1126/science.1121739)
3. Vijay Marupudi, Sashank Varma, Graded human sensitivity to geometric and topological concepts, Cognition, Volume 232, 2023, 105331, ISSN 0010-0277, <https://doi.org/10.1016/j.cognition.2022.105331>.
4. Hangbo Bao, Li Dong, Songhao Piao, Furu Wei (2021) BEiT: BERT Pre-Training of Image Transformers. DOI: <https://doi.org/10.48550/arXiv.2106.08254>
5. Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, Armand Joulin (2021) Emerging Properties in Self-Supervised Vision Transformers. DOI: <https://doi.org/10.48550/arXiv.2104.14294>
6. Zhengzhong Tu and Hossein Talebi and Han Zhang and Feng Yang and Peyman Milanfar and Alan Bovik and Yinxiao Li (2022) MaxViT: Multi-Axis Vision Transformer DOI: <https://arxiv.org/abs/2204.01697>
7. Zhuang Liu and Hanzi Mao and Chao-Yuan Wu and Christoph Feichtenhofer and Trevor Darrell and Saining Xie (2022) A ConvNet for the 2020s. DOI: <https://arxiv.org/abs/2201.03545>

Diff parameters for vision models

Lower layers of the models, only the final layer has the better results – forget about it

Different dataset, 1 model is enough, max\_vit on imagenet

Layers can be another aspect

Different training sets and checking on the model? Accuracy change?

Imagenet or tinyimagenet

Stretch goal - Om – finetuning the train model on geometry data