# Notes on the series of articles on the topic Deconstructing Neural Networks using Circuits

compiled by D.Gueorguiev, 5/18/2024

## Introductory notes

Given the popularity of the Neural Networks for Computer Vision it is clear that the scientific community needs *mechanistic interpretability* for those neural network architectures.

Speculative claims about Neural Networks operations and explainability of the latter

**Features**: *Features* are the fundamental unit of Neural Networks. They correspond to *Directions*. By *Directions* the authors mean linear combination of neurons in a layer. One can think of this as a direction vector in the vector space of activations of neurons in a given layer.

References

Zoom In: An Introduction to Circuits, Chris Olah et al, OpenAI, 2020: <https://distill.pub/2020/circuits/zoom-in>