Mechanistic Interpretability for Vision Models Optimization

23th July 2025

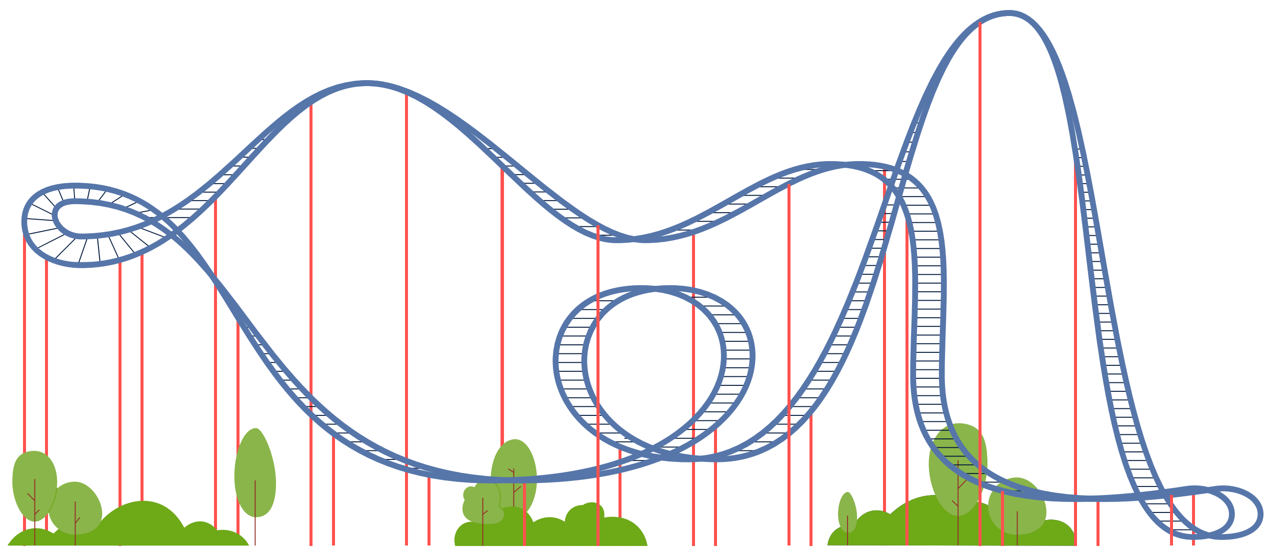
Computer Vision's course project

Authors:
Camilla Giuliani [1883207] & **Pietro D'Annibale** [1917211]

Project resources

https://colab.research.google.com/drive/1NNMyHI6ySeZPHcacPNtQd6y8-yUvGMZX#scrollTo=6jzzOI7xEby3

https://github.com/Sassotek/Mechanistic-Interpretability-for-Vision-Models-Optimization



A Note on the Journey

The development of this project has been a complex experience, much like a roller-coaster ride of ups and downs but each difficulty pushed us to grow and find alternative solutions.

Field of reasearch

- ViTs show very high performance on many vision task
- High computational cost makes ViTs not suitable for edge devices with limited hardware capabilities.

The project goal

Starting from a baseline ViT observe the model behavior and try to find a good compromise while trying to get a better inference time at the cost of a worse model accuracy cutting the edges in the computational graph that are less significative.

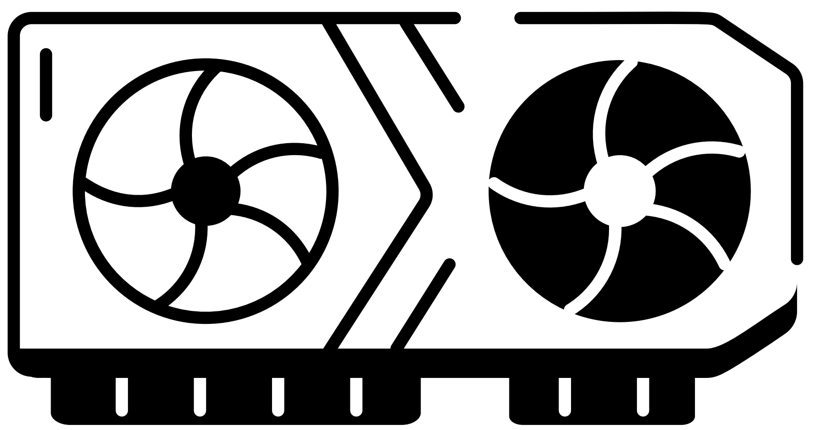
Overview

- Goal and ideas
- Hardware and settings
- Dataset
- Model Architecture
- Training
- ACDC → pruning
- Final results and evaluation
- References

Hardware & settings

2 different GPUs were used while working on colab

Nvidia GeForce RTX 3070



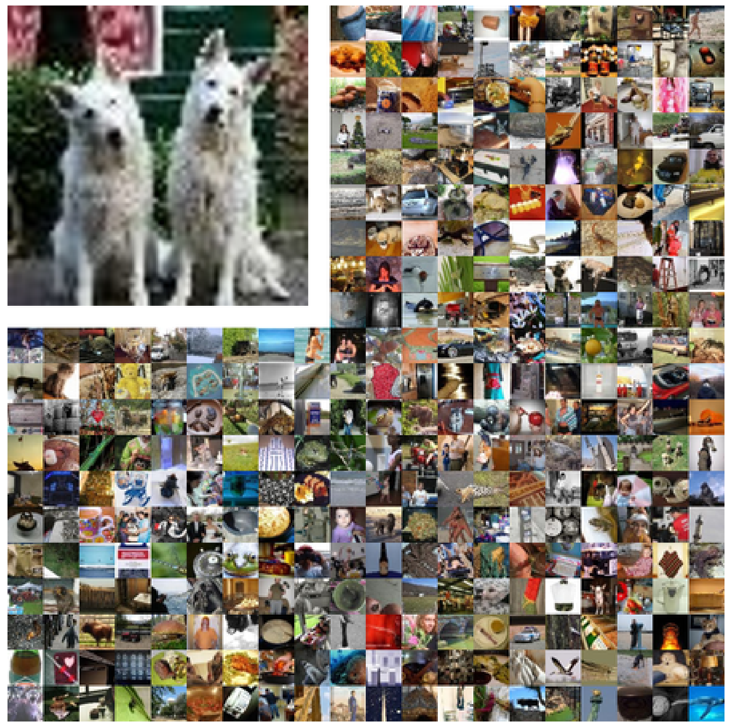
Nvidia Tesla T4

Dataset

Tiny ImageNet

- 200 classes
- 64x64 images

Augmentations applied {
  Random horizontal flip
  Random resized crop
  Random rotation
  Gaussian noise
  Random erasing
  Normalization with ImageNet mean & std
  CutMix
  MixUp
}



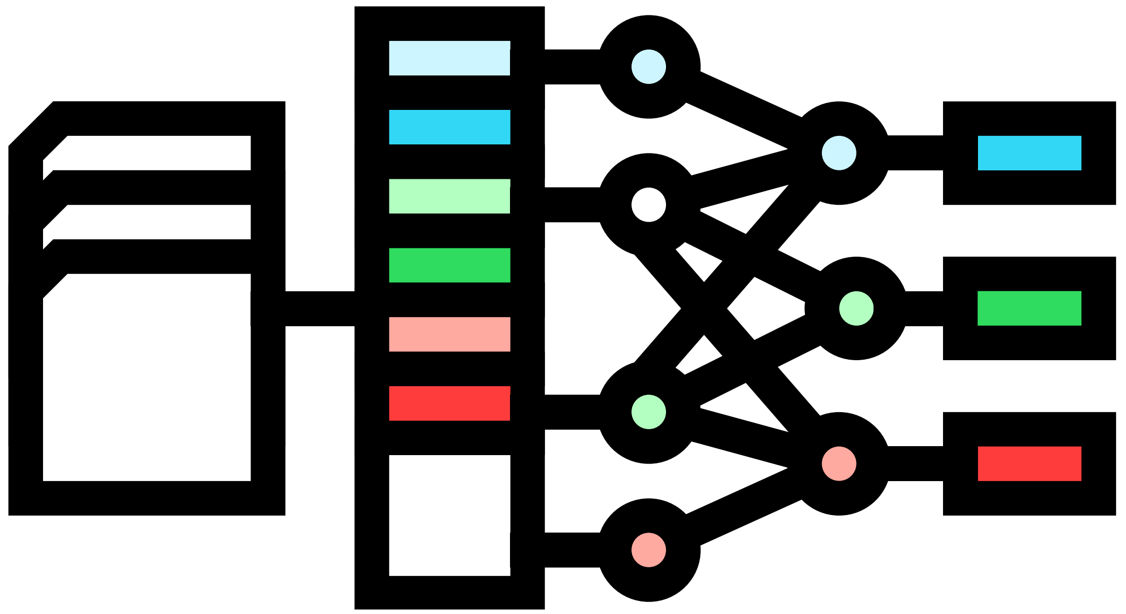
Model Architecture

We tested many different configurations of the Vit Model.

The last version is composed by

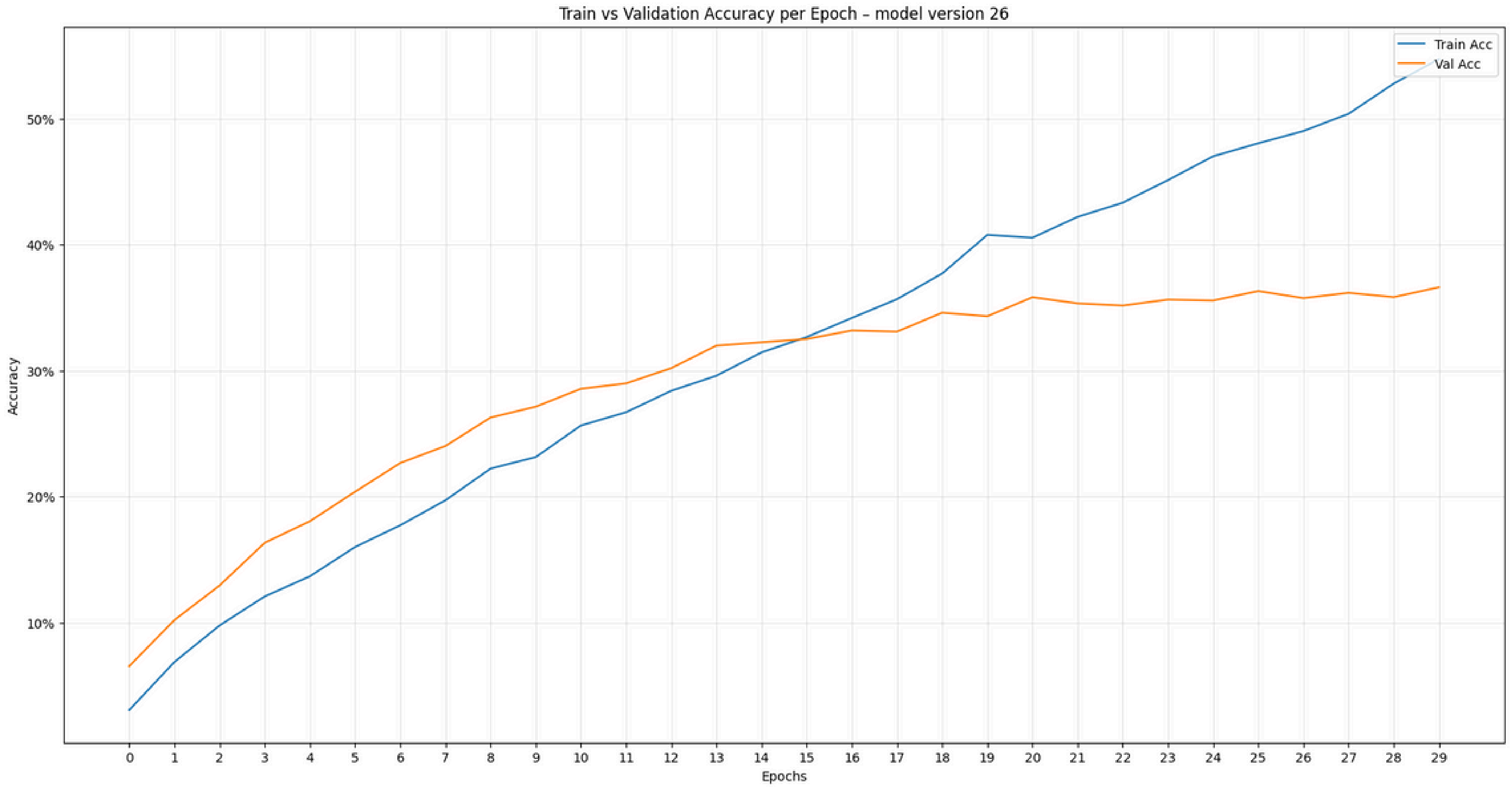
- Latent Size: 256
- Patch Size: 8
- 12 Encoders
- 8 MLP Heads

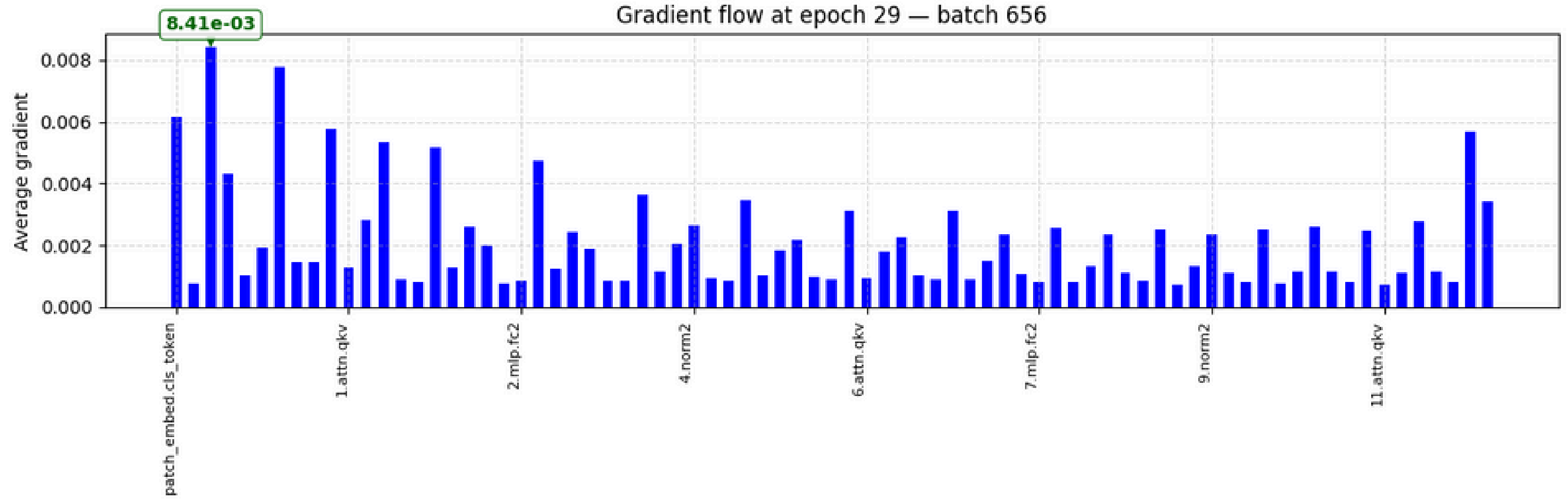
Other versions include bigger and smaller models, tested with different patch sizes.



Training

- 30 epochs
- Batch size: 128
- Gradient flow visualization
- AMP: Automatic Mixed Precision
- CosineAnnealingLR



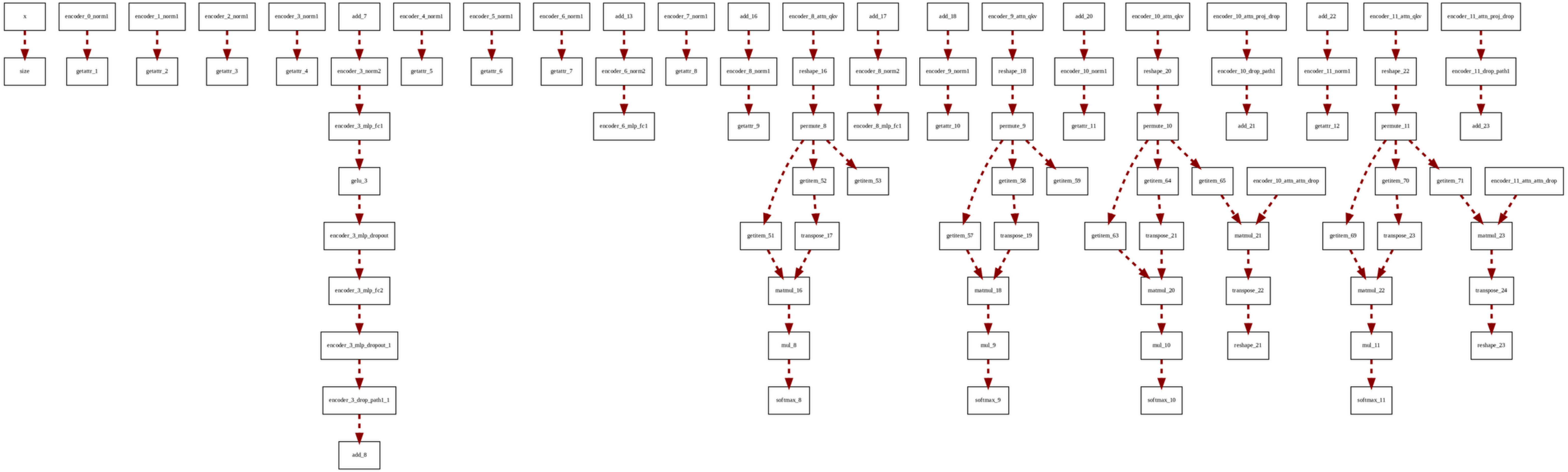


acdc

automatic circuit discovery

Pruning Phase

cuts = 81/518 edges



FINAL RESULTS

- Another 5 epochs training phase done on both pruned model and baseline model
- Not much changes in inference time
- The training gap between non-pruned and pruned model is recovered during training

	Model	Test Loss	Test Accuracy	InferenceTime
0	Baseline Model	361.098648	36.25%	70.08 s
1	Baseline Model Pruned	383.527886	33.00%	65.66 s
2	Upgraded Baseline Model	368.932721	36.34%	60.77 s
3	Upgraded Baseline Model Pruned	363.318928	36.38%	61.60 s

References

[1] A. Conmy et al. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability.
In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023)
[https://arxiv.org/abs/2304.14997]

[2] A. Syed, C. Rager and A.Conmy, (2024). Attribution Patching Outperforms Automated Circuit Discovery,
BlackboxNLP 2024.
[https://arxiv.org/abs/2310.10348]

[3] A. Vaswani et al. (2017). Attention is all you need.
In: Advances in Neural Information Processing Systems 36 (NeurIPS 2017)
[https://arxiv.org/abs/1706.03762]

[4] TinyImageNet dataset [https://www.kaggle.com/datasets/wissamsalam/tiny-imagenet-cleaned-for-classification]

[5] A.Dosovitskiy et al.(2021).An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
[https://arxiv.org/abs/2010.11929]

[6] Einops Guide [https://nbviewer.org/github/arogozhnikov/einops/blob/main/docs/1-einops-basics.ipynb]

[7] VISO.ai [https://viso.ai/deep-learning/vision-transformer-vit/]