Universität St.Gallen

Number of Attention Heads vs. Number of Transformer-Encoders in Computer Vision

Bernhard Bermeitinger, Tomas Hrycej, Siegfried Handschuh
KDIR2022, 2022-10-25, Valletta

Outline

1. Motivation
2. Parameter Structure of a Multi-Head Transformer
3. Measuring the Degree of Overdetermination
4. Experiment Results
    a. Used datasets
    b. Results
5. Conclusion

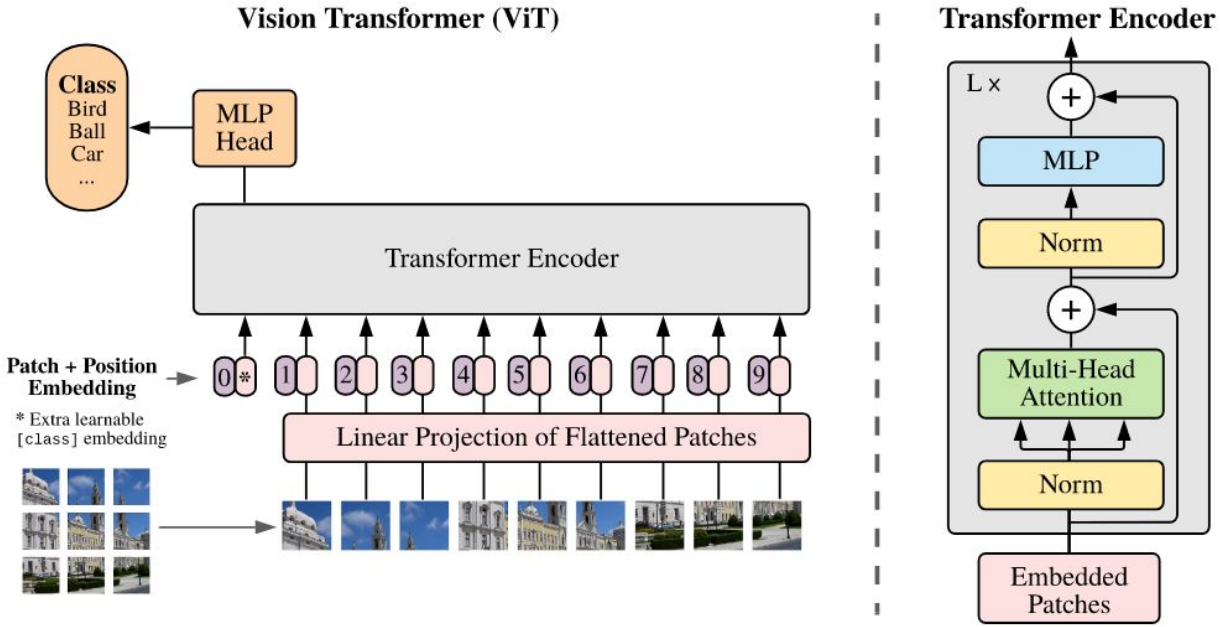
Motivation

- Transformers are widespread
  - Originally for Sequence-to-Sequence/Machine Translation
  - Self-Attention where each token *attends* each other token
- *Vision*-Transformer uses same concept

How many attention heads are necessary?

How do they interact when stacking multiple encoder layers?

Short recap: Vision Transformer



A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, Vienna, Austria, 2021, p. 21. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

Parameter Structure of Multiple Heads

- Multiple heads in one layer consist of:
  - Matrices transforming token vectors to a compressed form (*value*)
  - Matrices transforming token vectors to feature vectors (*key* and *query*)
  - Matrices transforming compressed and context-weighted tokens back to original token length
- Each encoder-layer has the same number of parameters
  - ⇒ parameter count proportional to number of encoders
- Adding more heads does not proportionally change parameter count
  - Number of weights in attention heads increase proportionally to attention heads
    But overall number of parameters not (due to MLP)

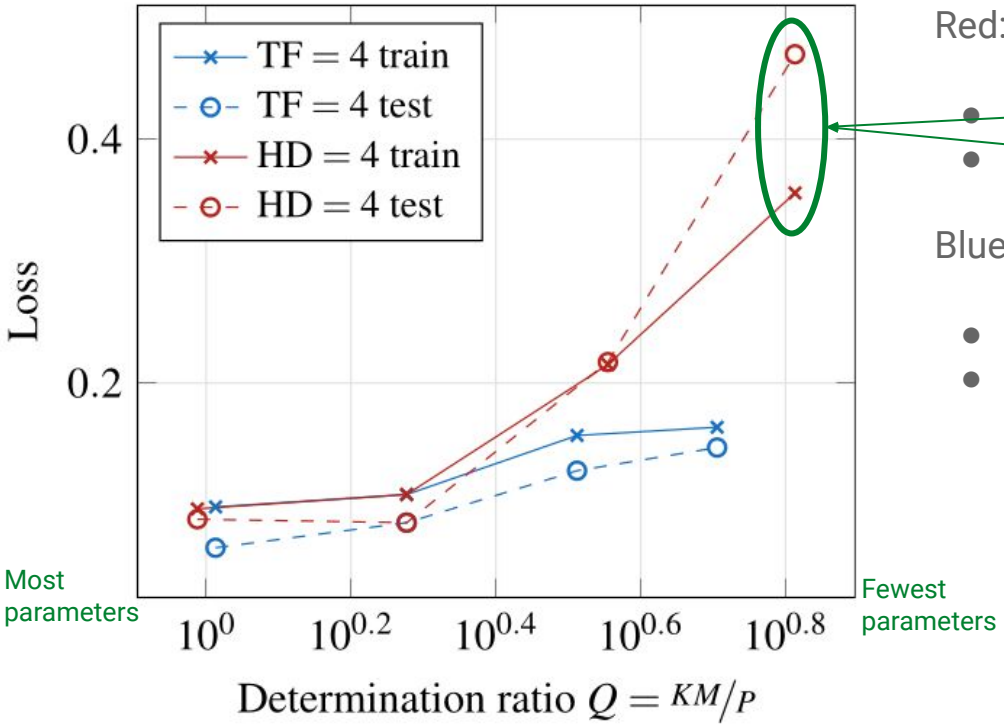
Measuring the Degree of Overdetermination

- Every fitting of input to output with a parameterized structure can be seen as an equation system.
  - Generally: *M* outputs with *K* training examples constitute *M*K* equations
  - The equations have *P* free parameters which are to be found by optimization
  - ⇒ *M*K* equations with *P* parameters
- Linear independent case: equations are satisfied by *M*K=P* parameters: The system is *exactly determined*.
- When P > M*K, the system has an infinite number of exact solutions: The system is *underdetermined*.
- When P < M*K, the system is *overdetermined*: no exact solution exists
- The ratio of overdetermination is given by: *Q = M*K / P*

Computing Results: Setup

- Various CV classification datasets are explored
- All experiments share the same setup:
  - Optimized by *AdamW*
  - Exclusively in *float32*
  - Batch size 256
  - Same data augmentation (translation, rotation, cropping, etc.)
- Images are scaled to a comparable size
- Patches are flattened
- Absolute positional encoding applied

Computing Result: MNIST



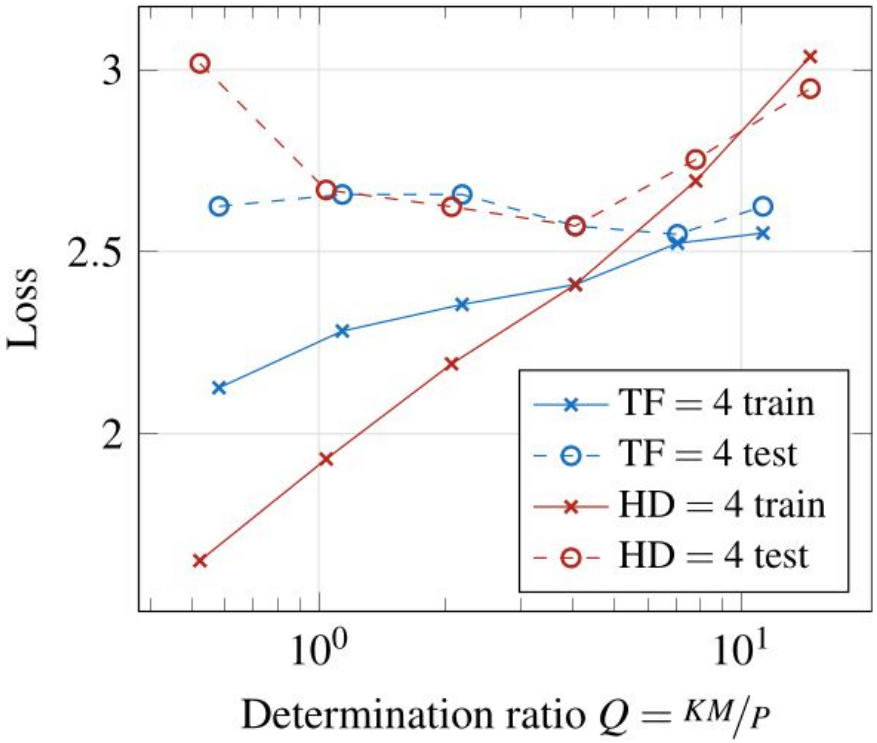
Red:

- 4 Attention Heads,
- 8, 4, 2, 1 Encoders

Blue:

- 4 Encoders
- 8, 4, 2, 1 Attention Heads

Computing Result: CIFAR-100



Red:

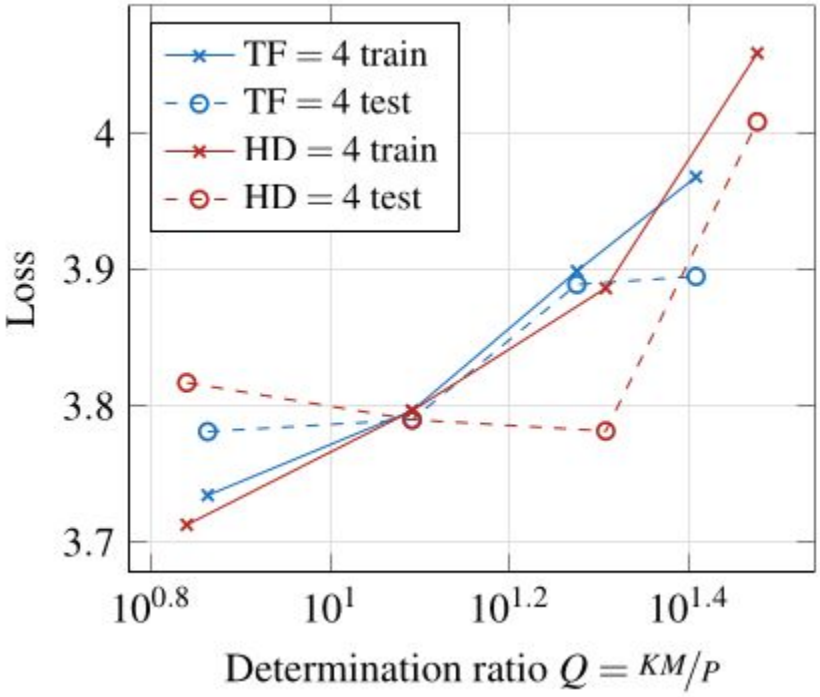
- 4 Attention Heads,
- 32, 16, 8, 4, 2, 1 Encoders

Blue:

- 4 Encoders
- 32, 16, 8, 4, 2, 1 Attention Heads



Computing Result: CUB-200-2011 (birds)



Red:

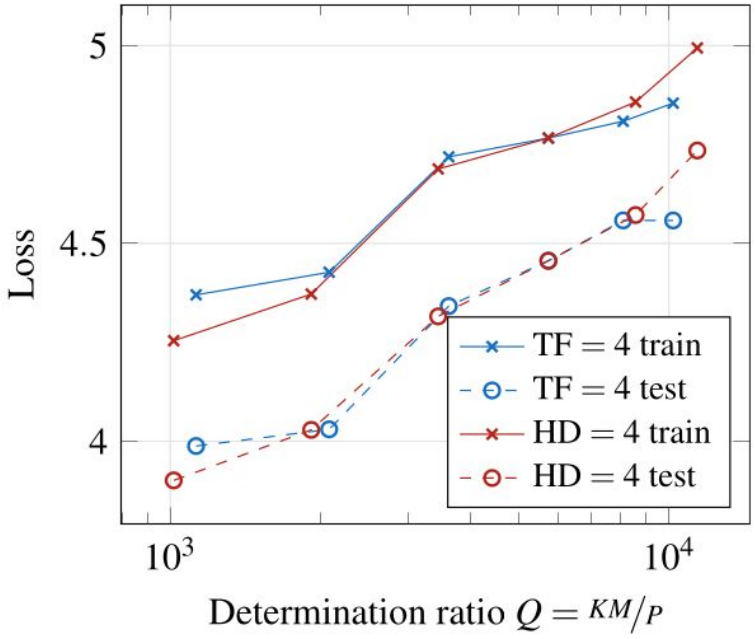
- 4 Attention Heads,
- 32, 16, 8, 4, 2, 1 Encoders

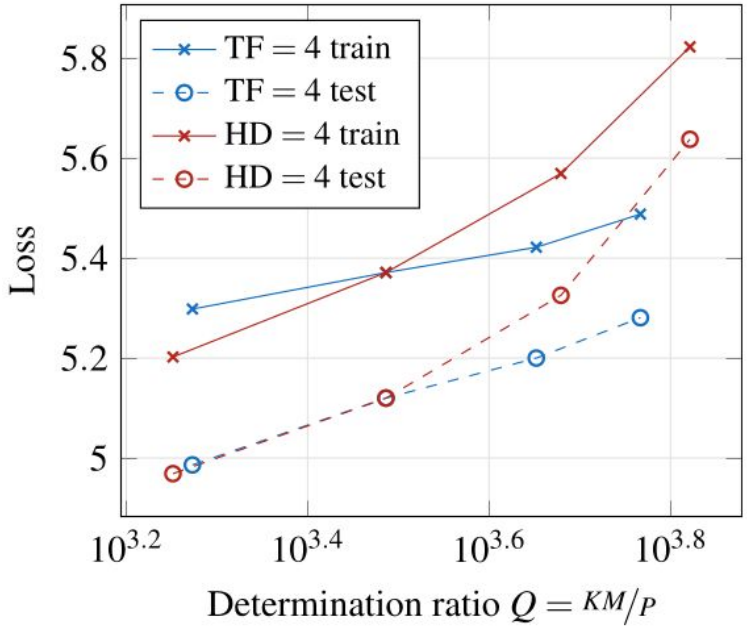
Blue:

- 4 Encoders
- 32, 16, 8, 4, 2, 1 Attention Heads



Results: Places365 and ImageNet





Conclusion

- Selecting the number of attention heads and number of transformer-encoders is an important choice.
- Determination ratio should exceed unity for similar train/test losses.

Generally, if role of context in the image is

- **Not important:** Invest in more parameters by increasing number of transformer-encoders. (low number of attention heads suffice)
- **Important:** Number of attention heads equally important as transformer-encoders