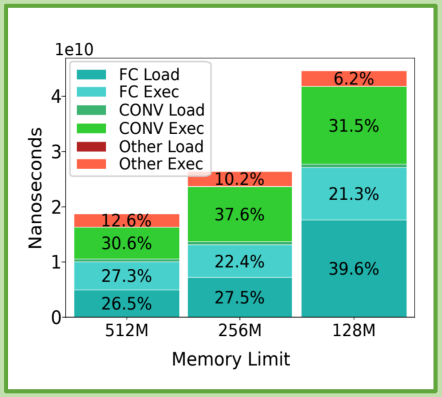
Modeling Inference Time of Deep Neural Networks on Memory-constrained Systems

Predicting the time needed to run a network based on a description of its layers and the amount of available memory.

1. Background

- Layer by layer execution (load, run, unload)
- Layers larger than available RAM
  - Requires swapping data to disk
  - Performance degrades
- Loading & execution time varies by layer type
  - 2 main categories: fully-connected & convolution
- Time taken per layer was measured under multiple levels of available memory and different hardware



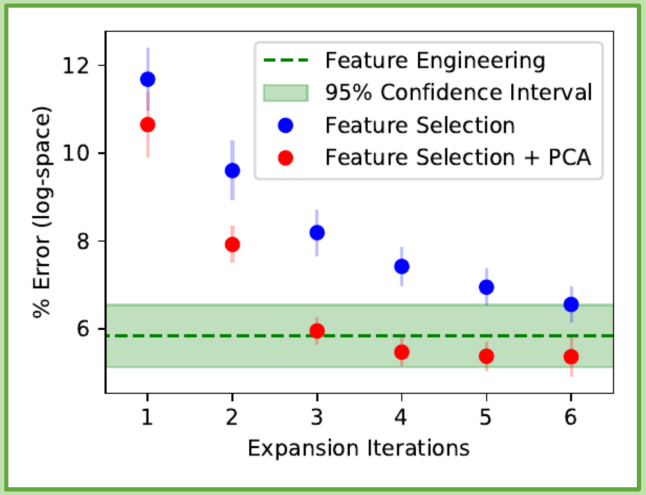
2. Approach

Two Models:

- **Feature Engineering**
  - Linear regression for each layer type
  - Predict based on theoretical calculations and amount of swapped memory
- **Feature Selection**
  - LASSO regression on all description features
  - Expand by multiplying features with each other
    - Adds non-linear relations between features
    - Applied iteratively
  - Principle component analysis
    - Reduces expanded features
    - Hopefully retains important predictive variance introduced by expansion

3. Results

- Models were trained on log-transformed data due to input spanning orders of magnitude
- In log-space, both approaches achieved ~6% mean absolute error
- However, after transformation predictions underestimated 2 – 10x
- Feature selection model seemed to overfit due to expanded data
  - Only used largest products
  - Features not consistently selected
- Models fit for different hardware were significantly different



4. Conclusions

- Feature engineering model
  - Simpler & straightforward interpretation
- Feature selection model
  - Automatically extensible with more data or more layer types
  - Models some non-linearity
- Future improvements
  - Model non-log-transformed data
  - Capture more non-linearity
  - Profile on more hardware

Hans Brouwer J.C.Brouwer@student.tudelft.nl

TU Delft