A Reduced Order Approach for Artificial Neural Networks Applied to Object Recognition

INTRODUCTION

Development pipeline of an Artificial Neural Network for the problem of Object Recognition to be later deployed in vision embedded systems.

A REDUCED ORDER APPROACH

Choose a dataset for the problem under consideration and determine the train dataset as the 80% of the complete dataset. Its components can be described as input-(expected) output pairs: \( \mathcal{D}_{\text{train}} = \{ (x(j), y(j)) \}_{j=1}^{N_{\text{train}}} \).

1. Original Neural Network
   - Take an Artificial Neural Network. It can be described as the compositions of its layers: \( \mathcal{A} \mathcal{N}^p = f_1 \circ f_{L-1} \circ \cdots \circ f_1 \).

2. Network Splitting
   - Choose a cut-off layer \( l \) (empirically). Define the pre- and post-models as follows: \( \mathcal{A} \mathcal{N}^p_{\text{pre}} = f_1 \circ f_{L-1} \circ \cdots \circ f_l \), \( \mathcal{A} \mathcal{N}^p_{\text{post}} = f_l \circ f_{L-2} \circ \cdots \circ f_1 \).

3. Dimensionality Reduction
   - Compute the pre-model outputs for each input in the train dataset.
   - Project them into a low-dimensional space of size \( R \), using the projection matrix obtained through the chosen reduction method.
   - Goal of the reduction methods: determine the most important parameters (eigenvalue analysis).

4. Predictor
   - The output of Step 3 becomes the input for the predictor, together with some additional inputs.
   - Maintain the predictor’s architecture of the original model, adjusting it for the variation in size of its inputs.

5. Compressed Neural Network
   - Re-training of the constructed Reduced Artificial Neural Network to achieve a good level of accuracy, comparable to the original model.

RESULTS

Some experimental results obtained with our reduction technique.

CONCLUSIONS

SOME FUTURE DEVELOPMENTS

- Generalizability of the approach
- More tasks, more architectures, more datasets

- Criteria for cut-off index
- Information theory notions (e.g. entropy) to understand the most important neurons/layers

- More reduction techniques
  - e.g. non linear ones

- Comparison and integration of other compression methods
  - e.g. pruning, quantization...

REFERENCES


CONTACT INFORMATION

- Laura Meneghetti, laura.meneghetti@sissa.it
- Nicola Demo, nicola.demo@fastcomputing.net
- Gianluigi Rozza, gianluigi.rozza@sissa.it

Check our GitHub page for the code!