Resource-constrained Neural Architecture Search on Language Models: A case study



INTRODUCTION

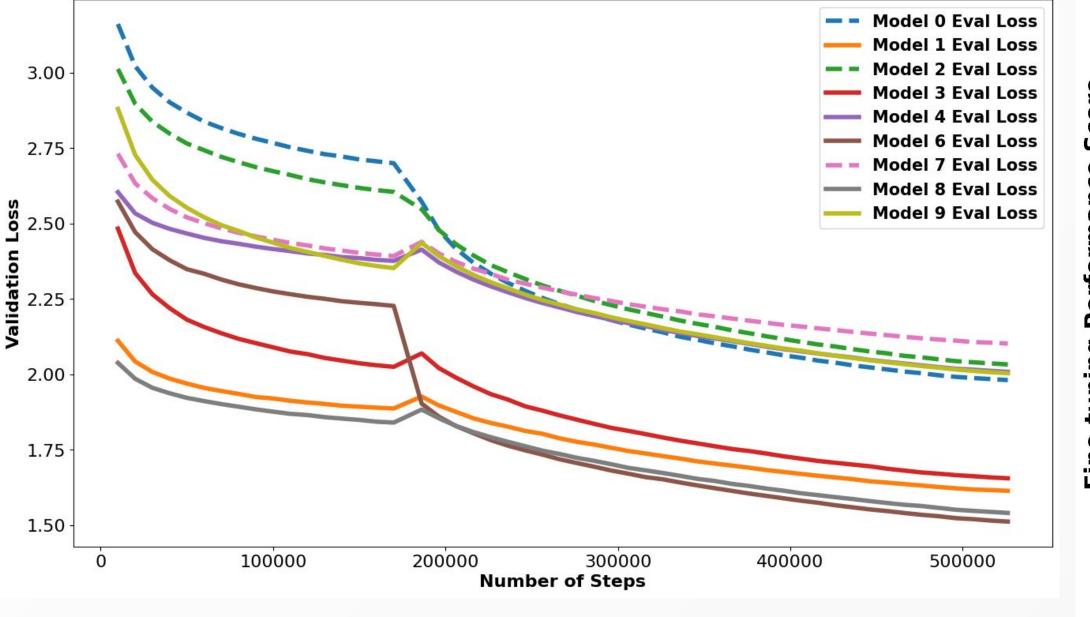
- Complex transformer architecture, limited intuition towards optimization
- Resource-constrained Neural Architecture Search (NAS) on the transformer-encoder macro-architecture (case study)

METHOD (PIPELINE) **Pareto Front Plot - Generation 1** next generation selection Pareto Fronts top offsprings • two-level hierarchical search space transformer encoder block (add/remove) offspring population parents population Feed-forward Neural Network (FNN) block (add/alter) original DistilBERT model Multi-head self-attention (alter) parent pre-training phase pre-training phase 0 selection mutation two-phase training with model 2. one round of 2. finetune downstream successive halving based task discarding on performance on MLM 3. define their fitness value

RESULTS

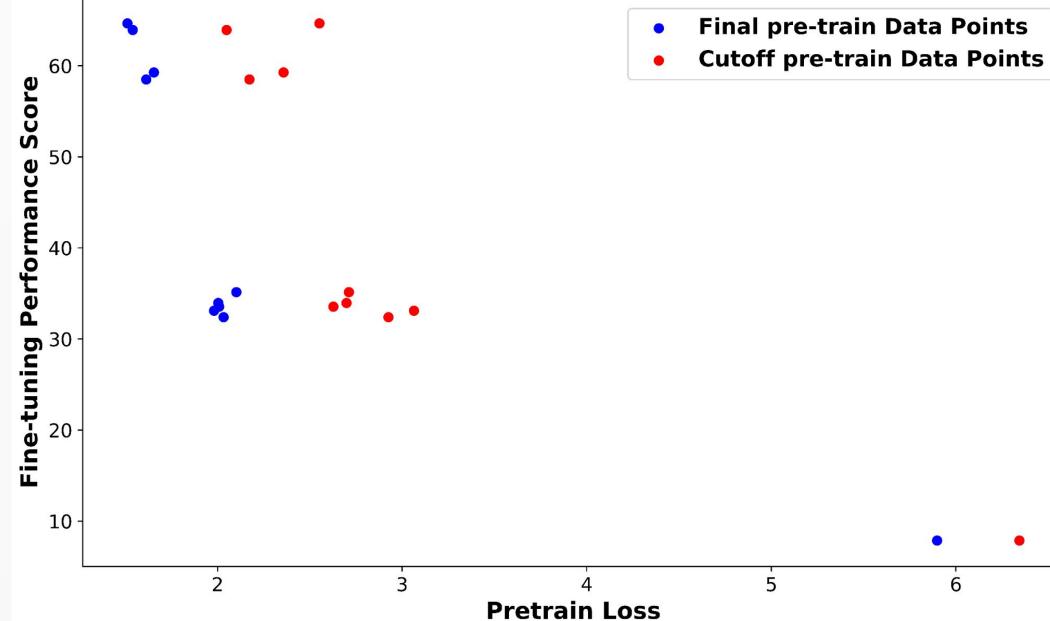
Multi-fidelity optimization:

minimal crossing



Cutoff criterion:

Strong correlation



DISCUSSION

- Strong Correlation between
 pre-training and downstream task
 performance
- Applicability of Learning Curve analysis and model selection mechanism

Challenges & Future work

- Search space definition
- LLM pre-training cost
 - weight inheritance
 - parameter/data/memory efficient pre-training

- Andreas Paraskeva
- João Pedro Reis
- Suzan Verberne
- Jan N. van Rijn

(Leiden University)(Universidade do Porto)(Leiden University)(Leiden University)





