Sparse Transformers are **(in)Efficient** Learners

Comparing Sparse Feedforward Layers in Small Transformers

Yijun Wu
Y.Wu-55@student.tudelft.nl

Maliheh Izadi
Professor

Arie van Deursen
Professor

Aral de Moor
Supervisor

1. Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The SparselyGated Mixture-of-Experts Layer, January 2017.

2. William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, June 2022.

3. Guillaume Lample, Alexandre Sablayrolles, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herve´ Jegou. Large Memory Layers with Product Keys, ´ December 2019.

1 Introduction

>60% of transformer parameters are in the feedforward layers. The first step to sample-efficient transformers is sample-efficient feedforward layers.

Sparse feedforward layers are the solution! Our research questions and findings are as follows:

Can sparse feedforward layers offer better language understanding than a dense feedforward network of the same size?

It's possible in the right configurations.

Are sparse feedforward layers faster than the feedforward network?

It's currently false for small models.

2 Sparse Feedforward Layers

Mixture of experts (MoE)

MoE[1] replaces the feedforward network with a gating network followed by many expert subnetworks. The gating network find the most compatible experts for the input. The MoE output is the linear combination of expert outputs weighed by their compatibility scores.

Controller Feedforward Network (CNT)

CNT[2] adds a learned controller to the standard feedforward network. The controller dynamically masks the activation vector, making it sparse.

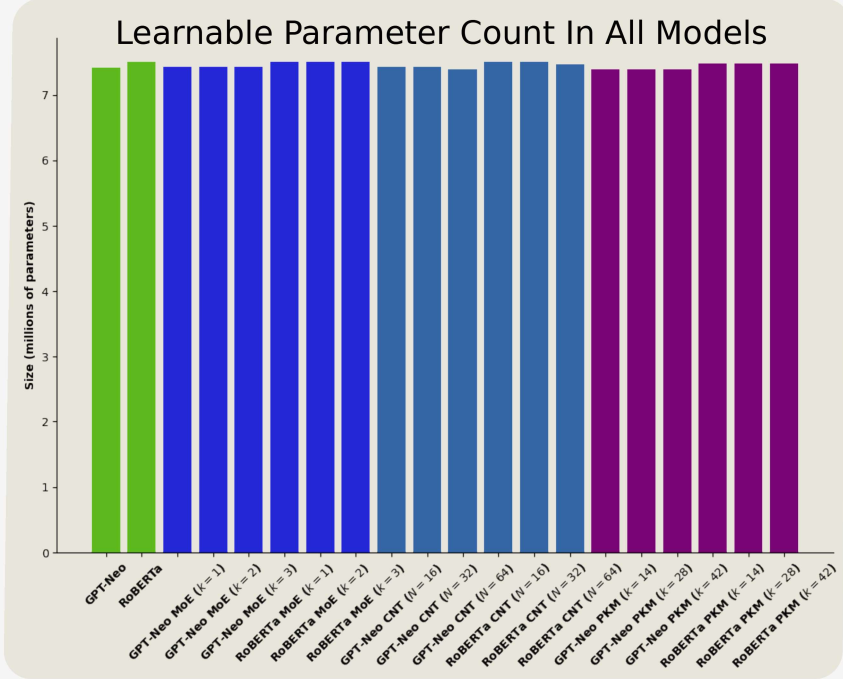
Product Key Memory (PKM)

PKM[3] consists of a query network, a key table, and a value table. The input is mapped to a query. The output is the linear combination of values weighed by the query-key similarity.

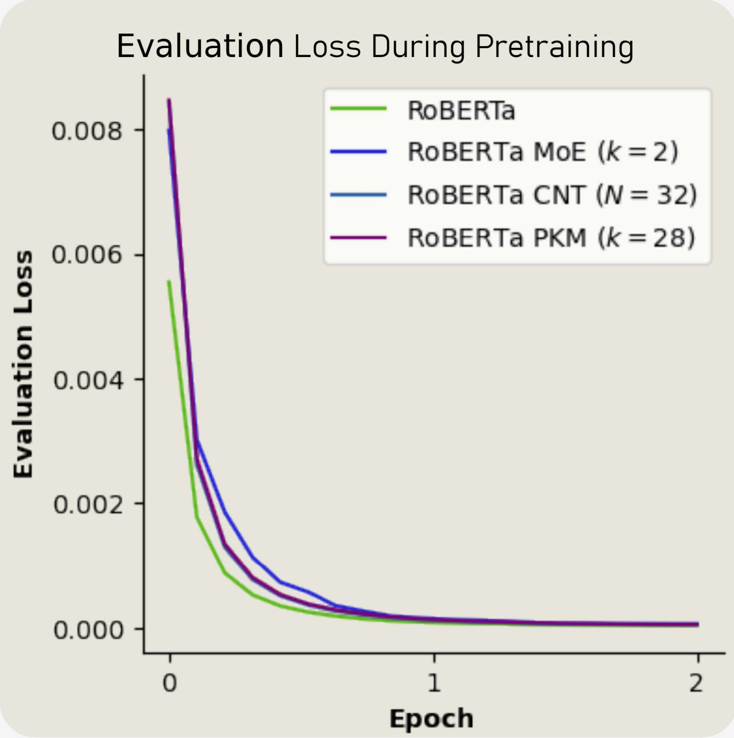
3 Method

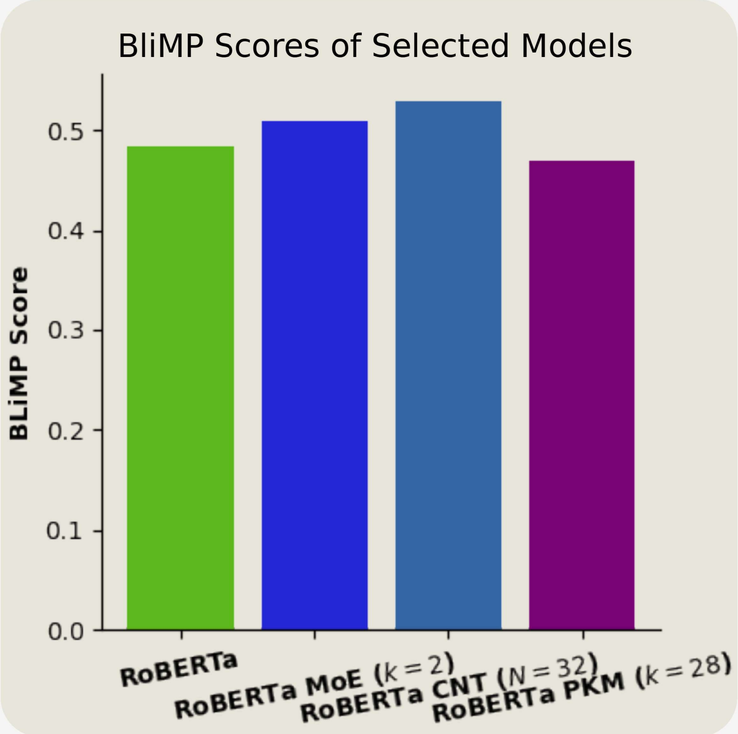
All our models are pretrained on the TinyStories dataset.

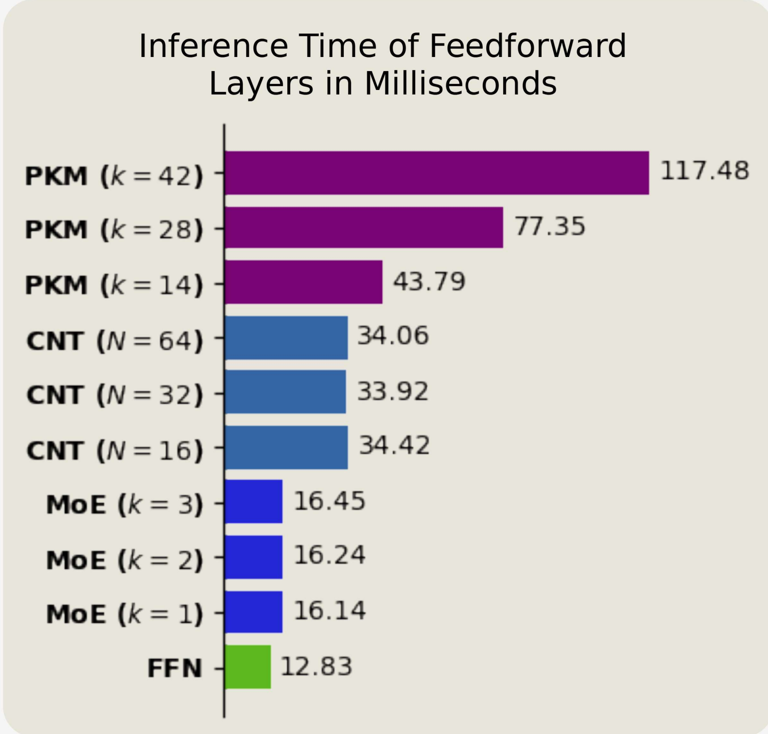
To test grammatical and language understanding we evaluate our models on BLiMP and (Super)GLUE tasks.



4 Results







5 Discussion & Future Work

Sparse models are flexible learners because they approximate larger networks than a feedforward network of the same size.

However small transformers on a single GPU do not enjoy the theoretical speed-ups sparsity and conditional computation bring.

Our research is limited by computational resources, which restricts our search space. We encourage future works in exploring sparse feedforward layers under more configurations.