Scaling Deep Neural Networks

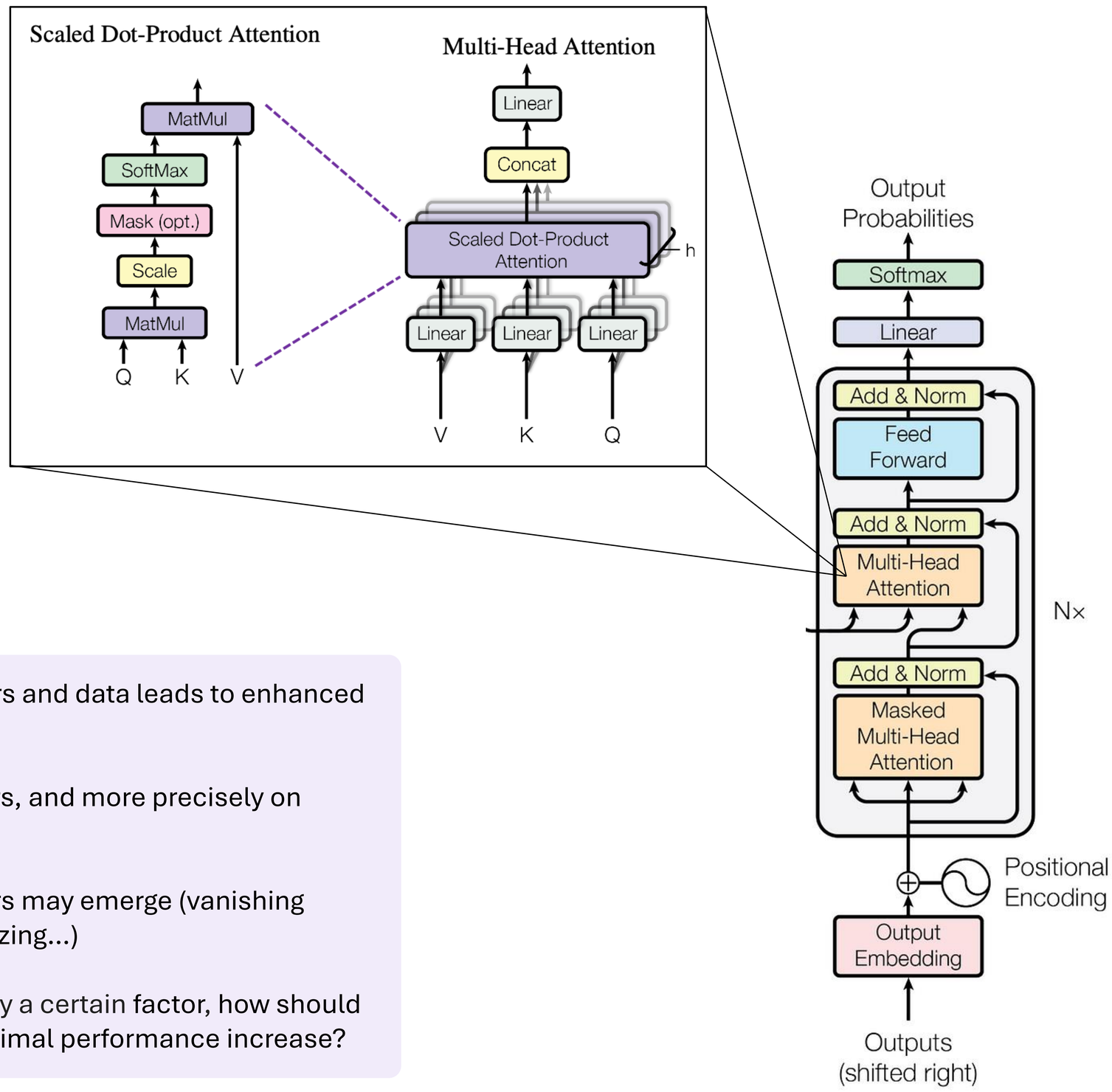
How To Tune Hyperparameters for a Transformer

Author: Chems Hamdi Supervisor: Lénaïc Chizat

Definitions

A transformer [1] is a deep learning model designed for sequential tasks such as natural language processing and computer vision. It uses self-attention to focus on relevant parts of the input, with **multi-head attention** parallelizing this process to capture diverse relationships.

Hyperparameters, such as the number of attention heads, layer dimensions, learning rate or initial distribution of weights, are settings that control the model's structure and training process, influencing its performance and capacity.



Decoder block of a transformer [1]

Motivations

Scaling Hypothesis: Increasing the model's parameters and data leads to enhanced performance.

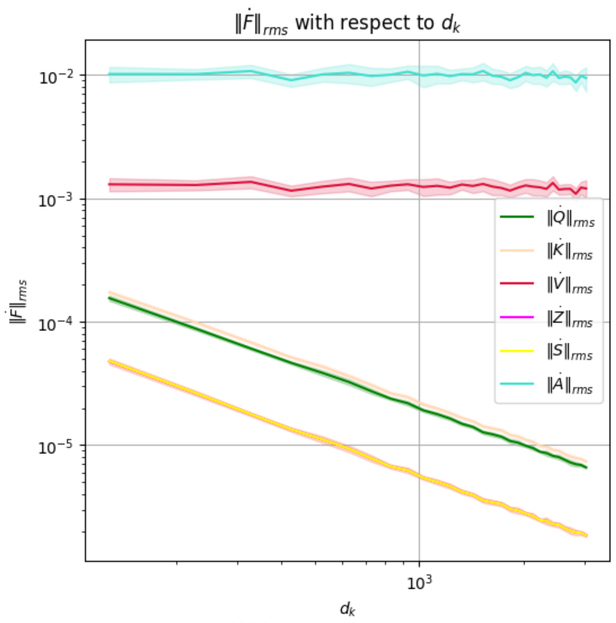
This study focuses on upscaling the model's parameters, and more precisely on scaling its dimensions (size).

Context: as model size increases undesirable behaviors may emerge (vanishing gradients, exploding gradients, loss spikes, feature freezing...)

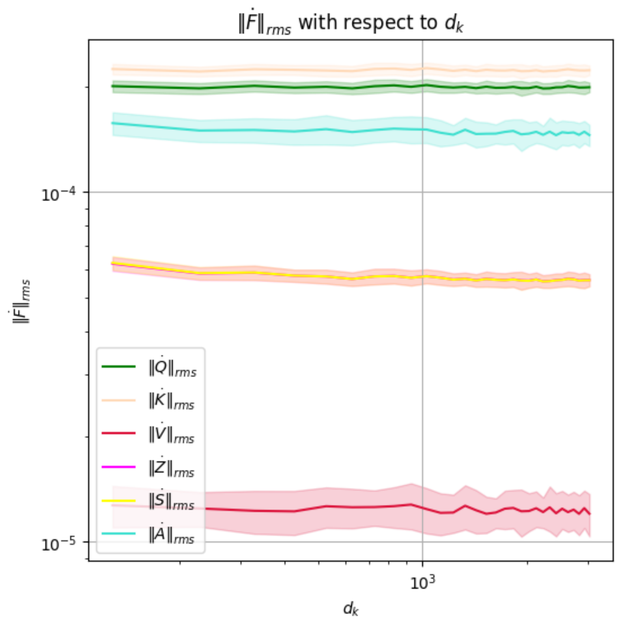
Research Question: If we scale the size of the model by a certain factor, how should we change the hyperparameters so as to obtain the optimal performance increase?

Effect

The two figures present the effect of **tuning hyperparameters** such that the process of **learning features becomes asymptotically stable** with dimension dk. Each line correspond to an intermediate layer in the block.



Without tuned parameters



With tuned parameters

Method

- We look at the asymptotic behavior of each intermediate mapping of the Forward and Backward Pass at initialization
- Relying on the **Feature Speed Formula [2]** and a **list of desirable properties**, we derive **hyperparameters scaling rules**, to ensure optimal training dynamics

Results

- We show that certain dimensions cannot be scaled while satisfying the so-called **"FSC" desiderata**
- For the other dimensions, we propose new HP scaling rules

Access full report

References

[1] A. Vaswani *et al.*, « Attention Is All You Need », 1 août 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[2] L. Chizat et P. Netrapalli, « The Feature Speed Formula: a flexible approach to scale hyper-parameters of deep neural networks », 22 juin 2024, *arXiv*: arXiv:2311.18718. doi: 10.48550/arXiv.2311.18718.

LAIDLAW SCHOLARS EPFL