Your Transformer is Secretly Linear

8th June 2024

Transformers

1. Large number of parameters →Subject to overfitting
2. As size grows → More hardware required for inference (bottleneck)
3. Hard to visualize/understand all the layers and importance (as opposed to CNN's where several hierarchical features can be understood)

Efficient transformers

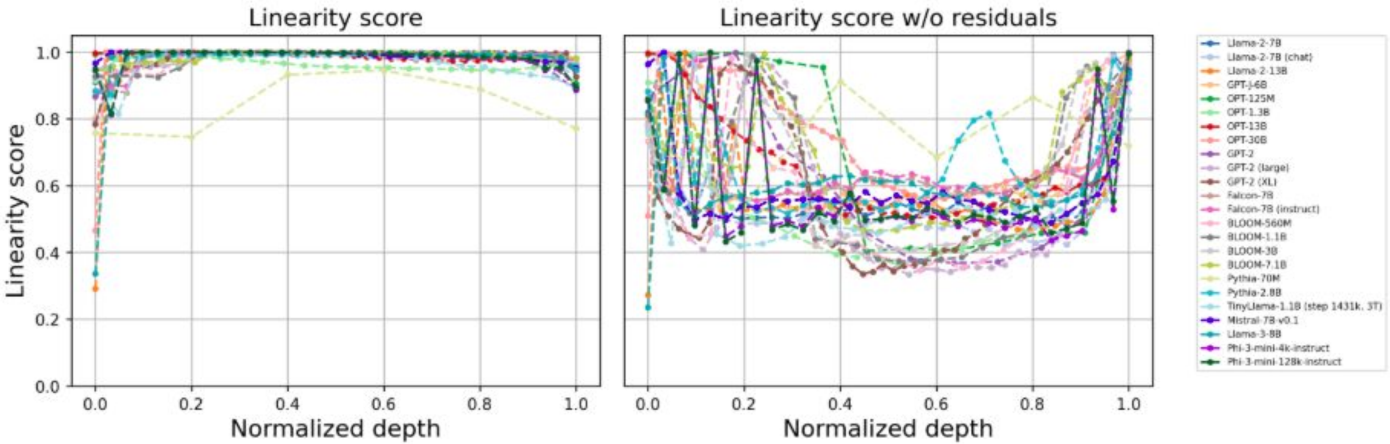
1. https://arxiv.org/pdf/2009.06732 (Efficient Transformers: A Survey)

Key concepts

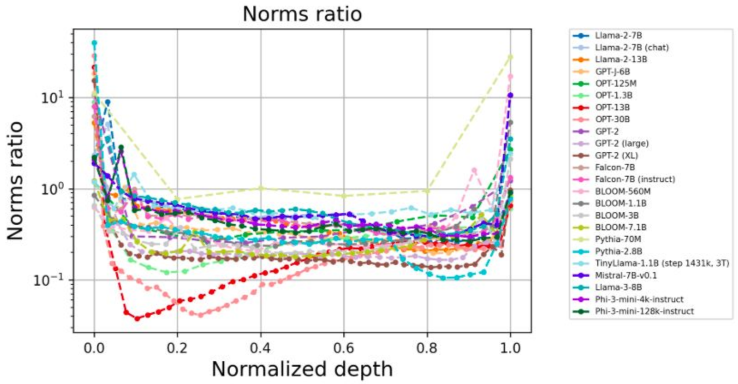
1. Embedding transformations can be thought of as change of the values of input embeddings as they go through the different transformer layers
2. Residual component refers to the skip connections that are present in transformers i.e. Output = f(input)+Input
    a. Introduced originally to solve the problem of vanishing gradients
    b. Keep the input context intact
3. Output norm is the magnitude of the output from a layer
4. Linearity score measured by Procrustes similarity

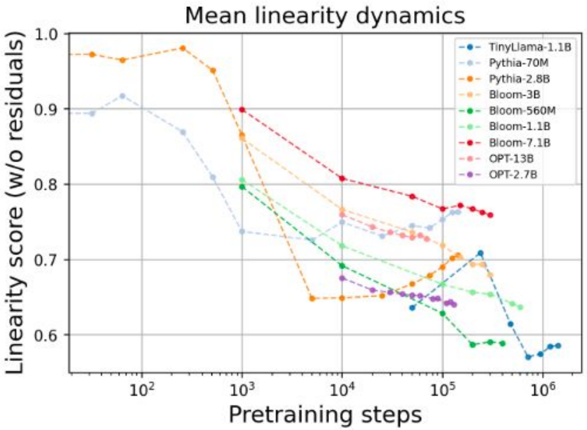
Main motivation (make the model smaller)

1. Understand and quantify this linearity
2. Depth pruning
3. Cosine-similarity-based regularization, aimed at reducing layer linearity (having more information / layer)



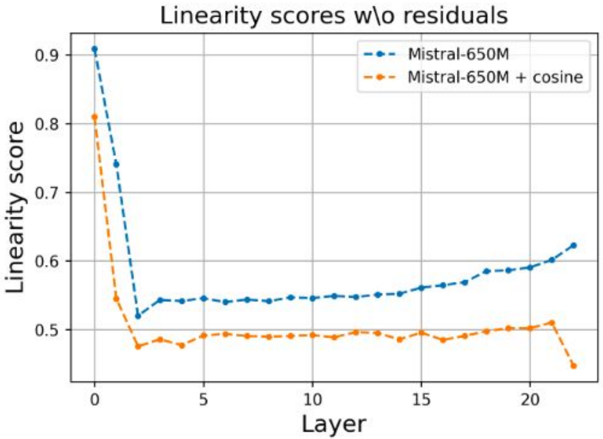
1. Normalized depth → Layer index / Total number of layers
2. Linearity score w/o residuals is non-linear since the transformation function can be highly non-linear.. Output = f(input), where as with residuals Output = f(input)+Input the linearity score is ~0.99
3. Norms ratio = f(input) / Output

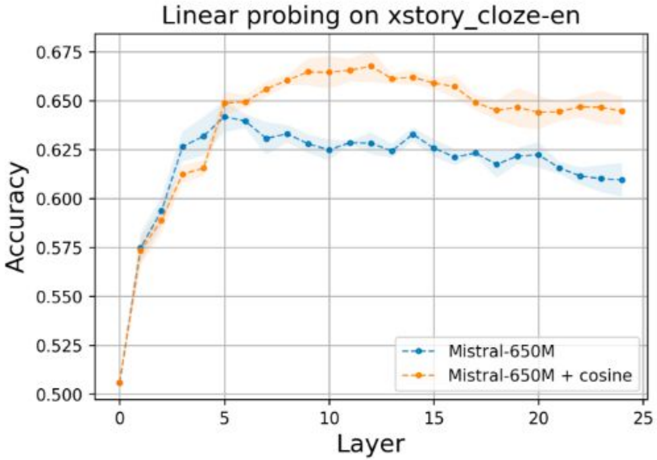




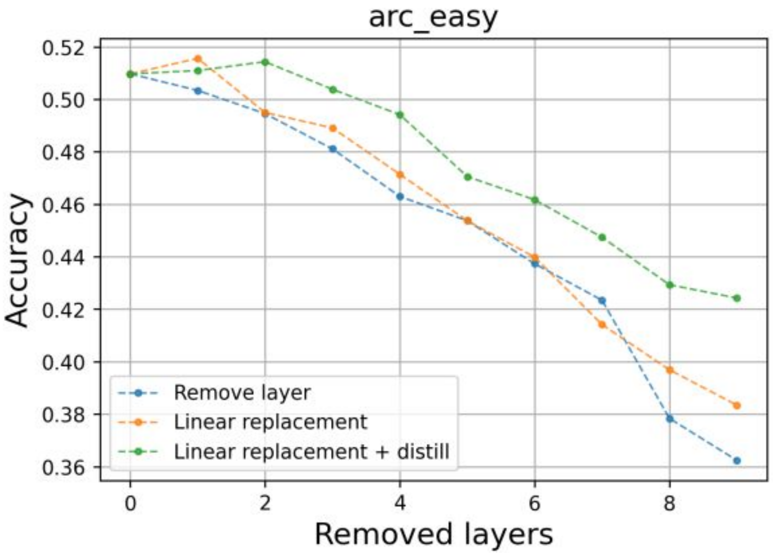
Model Name	Super_Glue/MultiRC	Super_Glue/BoolQ	Super_Glue/CB	Reward Modeling
OPT-125M	0.085 ± 0.008	0.217 ± 0.038	0.048 ± 0.009	0.060 ± 0.008
OPT-1.3B	0.055 ± 0.021	0.382 ± 0.004	0.088 ± 0.010	0.062 ± 0.007
OPT-2.7B	0.061 ± 0.025	0.356 ± 0.005	0.066 ± 0.029	0.054 ± 0.003
Llama2-7B	0.141 ± 0.006	0.051 ± 0.024	0.081 ± 0.070	0.194 ± 0.027
GPT2	0.085 ± 0.021	0.048 ± 0.016	0.004 ± 0.003	0.092 ± 0.013
GPT2-Large	0.049 ± 0.003	0.023 ± 0.008	0.025 ± 0.014	0.085 ± 0.008
GPT2-XL	0.040 ± 0.007	0.037 ± 0.007	0.028 ± 0.019	0.038 ± 0.008

Table 1: Delta of linearity score w/o residuals after fine-tuning various tasks. Note that all values are strictly positive, which means that linearity always increases during fine-tuning.





XStoryCloze consists of the professionally translated version of the English StoryCloze dataset (Spring 2016 version) to 10 non-English languages. This dataset is intended to be used for evaluating the zero- and few-shot learning capabilities of multlingual language models. This dataset is released by Meta AI.


arc_easy
Accuracy
0.52
0.50
0.48
0.46
0.44
0.42
0.40
0.38
0.36
Removed layers
0
2
4
6
8
Remove layer
Linear replacement
Linear replacement + distill