LLaMA: Open Efficient Foundation Language Models

23.06.30

이정민

LLaMA

Paper

❖ LLaMA: Open and Efficient Foundation Language Models

- [2023, Computation and Language] 23.06.29 기준 506회 인용

LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave*, Guillaume Lample*

Meta AI

Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community[1].

performance, a smaller one trained longer will ultimately be cheaper at inference. For instance, although Hoffmann et al. (2022) recommends training a 10B model on 200B tokens, we find that the performance of a 7B model continues to improve even after 1T tokens.

The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. The resulting models, called *LLaMA*, ranges from 7B to 65B parameters with competitive performance

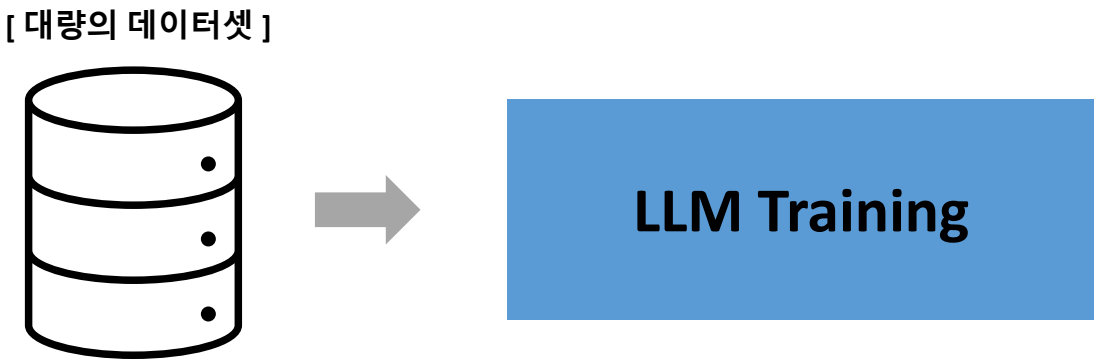
Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

LLaMA

LLaMA: Open and Efficient Foundation Language Models

❖ 연구 배경

- 제한된 computing budget 내에서 최고의 성능은 모델을 더 키울 때가 아니라 **작은 모델을 보다 많은 데이터를 이용해서 학습** 시킬 때 달성됨(by Chinchilla)

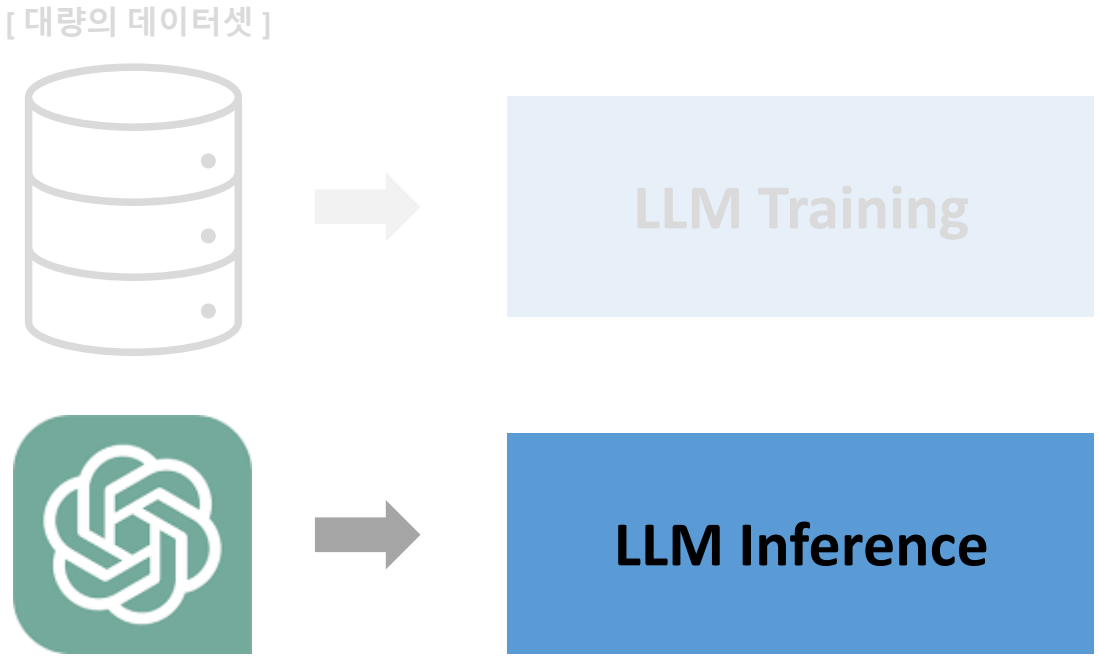


LLaMA

LLaMA: Open and Efficient Foundation Language Models

❖ 연구 배경

- 제한된 computing budget 내에서 최고의 성능은 모델을 더 키울 때가 아니라 **작은 모델을 보다 많은 데이터를 이용해서 학습** 시킬 때 달성됨(by Chinchilla)
- Chinchilla는 실제 LLM 서비스화에 중요한 **inference budget** 간과



LLaMA

LLaMA: Open and Efficient Foundation Language Models

❖ 기여점

- 주어진 inference budget 내에서 최고의 성능을 달성할 수 있는 LM 모델들을 더 많은 tokens를 사용해서 학습
- Open-source로 사용 가능하도록 **publicly available data**만 사용해서 학습

LLaMA

LLaMA: Open and Efficient Foundation Language Models

❖ Architecture based Transformer

- Pre-Normalization (GPT-3)
  - ➢ **학습 안정성**을 위해 transformer sub-layer의 output이 아닌 input에 RMSNorm 사용

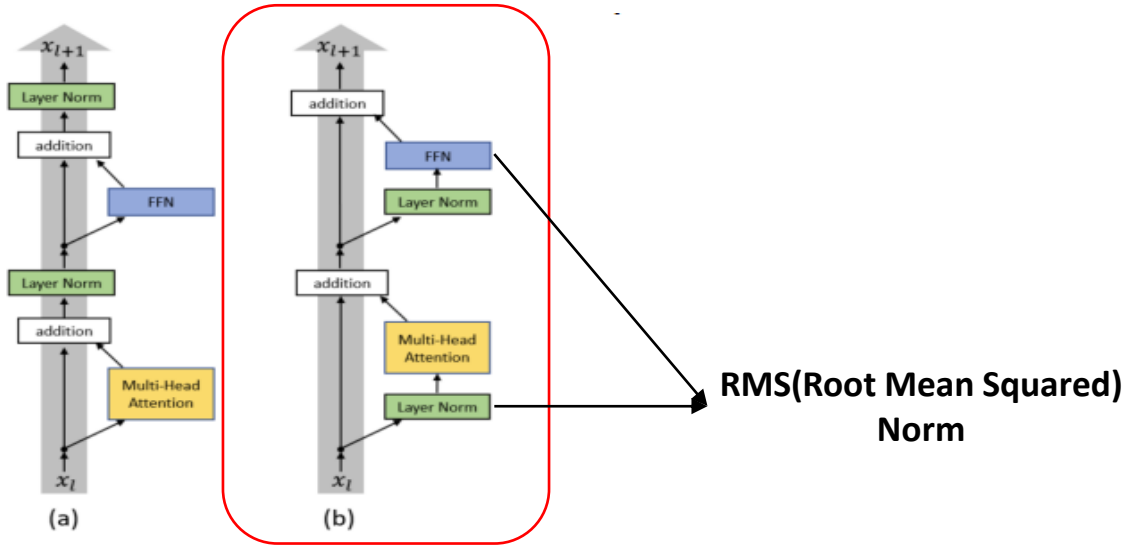


Figure 1. (a) Post-LN Transformer layer; (b) Pre-LN Transformer layer.

[Pre-Normalization]

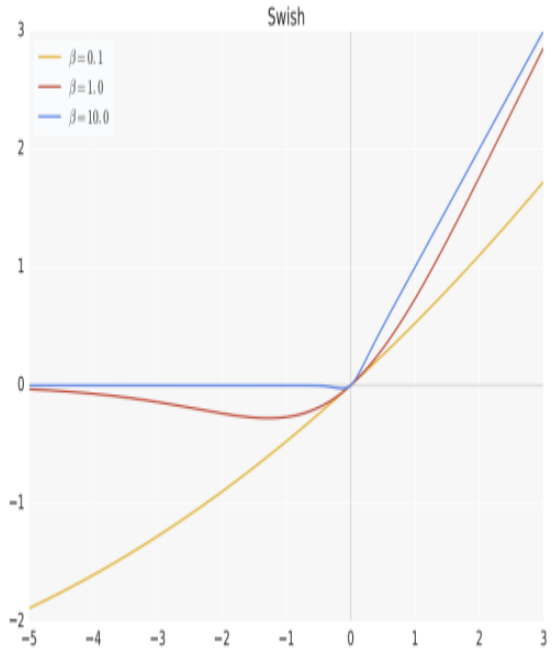
Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., ... & Liu, T. (2020, November). On layer normalization in the transformer architecture. In *International Conference on Machine Learning* (pp. 10524-10533). PMLR.

LLaMA

LLaMA: Open and Efficient Foundation Language Models

❖ Architecture based Transformer

- SwiGLU activation function (PaLM)



[SwiGLU activation]

$$Swish_\beta(x) = x\sigma(\beta x)$$

$$GLU(x, W, V, b, c) = \sigma(xW + b) \otimes (xV + c)$$

$$\boldsymbol{SwiGLU(x, W, V) = xW\sigma(xW) \otimes xV}$$

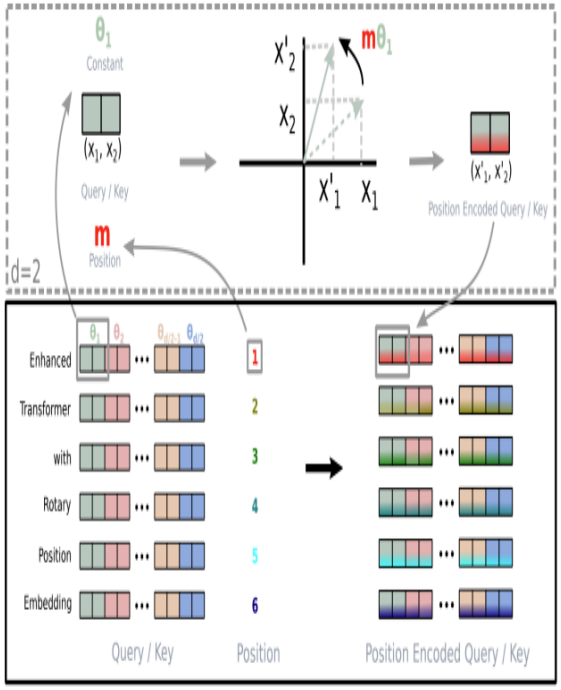
- ❖ $Swish$는 $ReLU$의 장점을 가지면서 미분 가능하며 단조 증가 함수가 아니라는 점에서 추가적인 장점을 가짐
- ❖ β : Hyper parameter
- ❖ W, V: 학습 대상

LLaMA

LLaMA: Open and Efficient Foundation Language Models

❖ Architecture based Transformer

- Rotary Embeddings (GPTNeo)



[Rotary Embeddings]

$$\theta_i = 10000^{-2i/d}$$

$$R^d_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix}$$

❖ Long-term decay

- Relative Position이 멀어지면 Inner product 값이 감소
- 먼 거리는 적은 유사도를 줌

❖ Rope with Linear attention

- 보다 효율적인 연산 가능

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2021). Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

LLaMA

LLaMA: Open and Efficient Foundation Language Models

❖ Efficient Implementation

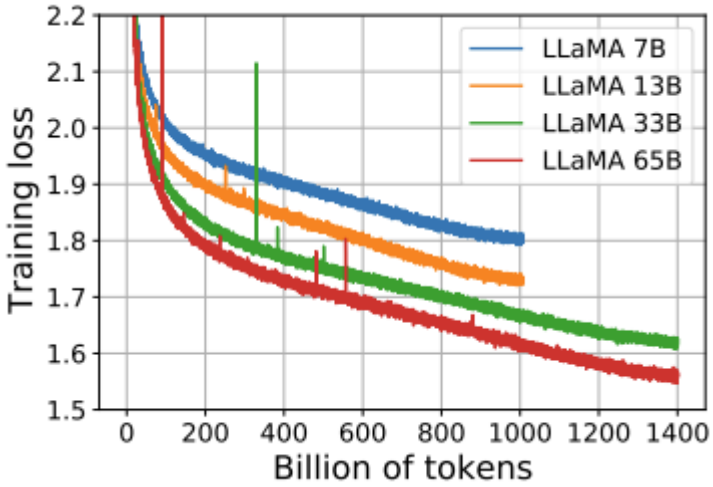
- Efficient implementation of the causal multi-head attention operator 사용
  - ➢ Attention weights 저장 x, masked tokens에 대한 key/query score 계산 x
- Linear layer 후 계산되는 activation 값 저장 및 재사용
- 2,048개의 A100(80GB) GPU를 사용하여 1.4T tokens를 학습하기 위해 약 21일 소요

LLaMA

LLaMA: Open and Efficient Foundation Language Models

❖ Experiments

- 모든 크기의 모델에서 더 많은 tokens를 학습시켜도 성능이 지속적으로 향상
- Zero-shot setting에서 우수한 성능을 도출함



		BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
GPT-3	175B	60.5	81.0	-	78.9	70.2	68.8	51.4	57.6
Gopher	280B	79.3	81.8	50.6	79.2	70.1	-	-	-
Chinchilla	70B	83.7	81.8	51.3	80.8	74.9	-	-	-
PaLM	62B	84.8	80.5	-	79.7	77.0	75.2	52.5	50.4
PaLM-cont	62B	83.9	81.4	-	80.6	77.0	-	-	-
PaLM	540B	**88.0**	82.3	-	83.4	**81.1**	76.6	53.0	53.4
LLaMA	7B	76.5	79.8	48.9	76.1	70.1	72.8	47.6	57.2
	13B	78.1	80.1	50.4	79.2	73.0	74.8	52.7	56.4
	33B	83.1	82.3	50.4	82.8	76.0	**80.0**	**57.8**	58.6
	65B	85.3	**82.8**	**52.3**	**84.2**	77.0	78.9	56.0	**60.2**

Table 3: **Zero-shot performance on Common Sense Reasoning tasks.**

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

LLaMA

LLaMA: Open and Efficient Foundation Language Models

❖ Experiments

- 여러 few-shot setting 및 여러 task에서 우수한 성능을 도출함

		0-shot	1-shot	5-shot	64-shot
GPT-3	175B	14.6	23.0	-	29.9
Gopher	280B	10.1	-	24.5	28.2
Chinchilla	70B	16.6	-	31.5	35.5
PaLM	8B	8.4	10.6	-	14.6
	62B	18.1	26.5	-	27.6
	540B	21.2	29.3	-	39.6
LLaMA	7B	16.8	18.7	22.0	26.1
	13B	20.1	23.4	28.1	31.9
	33B	**24.9**	28.3	32.9	36.0
	65B	23.8	**31.0**	**35.0**	**39.9**

Table 4: **NaturalQuestions.** Exact match performance.

		0-shot	1-shot	5-shot	64-shot
Gopher	280B	43.5	-	57.0	57.2
Chinchilla	70B	55.4	-	64.1	64.6
LLaMA	7B	50.0	53.4	56.3	57.6
	13B	56.6	60.5	63.1	64.0
	33B	65.1	67.9	69.9	70.4
	65B	**68.2**	**71.6**	**72.6**	**73.0**

Table 5: **TriviaQA.** Zero-shot and few-shot exact match performance on the filtered dev set.

		RACE-middle	RACE-high
GPT-3	175B	58.4	45.5
PaLM	8B	57.9	42.3
	62B	64.3	47.5
	540B	**68.1**	49.1
LLaMA	7B	61.1	46.9
	13B	61.6	47.2
	33B	64.1	48.3
	65B	67.9	**51.6**

Table 6: **Reading Comprehension.** Zero-shot accuracy.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.