Large Language Models

Questions

Flavio Giobergia

For the following decoder-only transformer model, compute:

1) the number of parameters for the embedding layer
2) The number of parameters of the transformer (excluding the embedding layer, or head layer(s))

Note: assume that none of the layers uses a bias term

- Vocabulary size V=30,000
- Embedding/model dimension d=768
- Positional embeddings: learned, for 512 positions
- Number of layers L=12
- Number of attention heads H=12
- FF-NN projection dimension: 3072
- Head dimension $d_k = d_v = 64$

- Vocabulary size V=30,000
- Embedding/model dimension d=768
- Positional embeddings: learned, for 512 positions
- Number of layers L=12
- Number of attention heads H=12
- FF-NN projection dimension: 3072
- Head dimension $d_k = d_v = 64$

parameters (Embedding)

- (Token embedding size + positional embedding size)
- 30000 * 768 + 512 * 768 = 23,433,216

parameters (transformer)

12*(attention + FF)

Attention:

Wq, Wv, Wk, Wo = (768*64*12)*3

Wo = 64*12*768

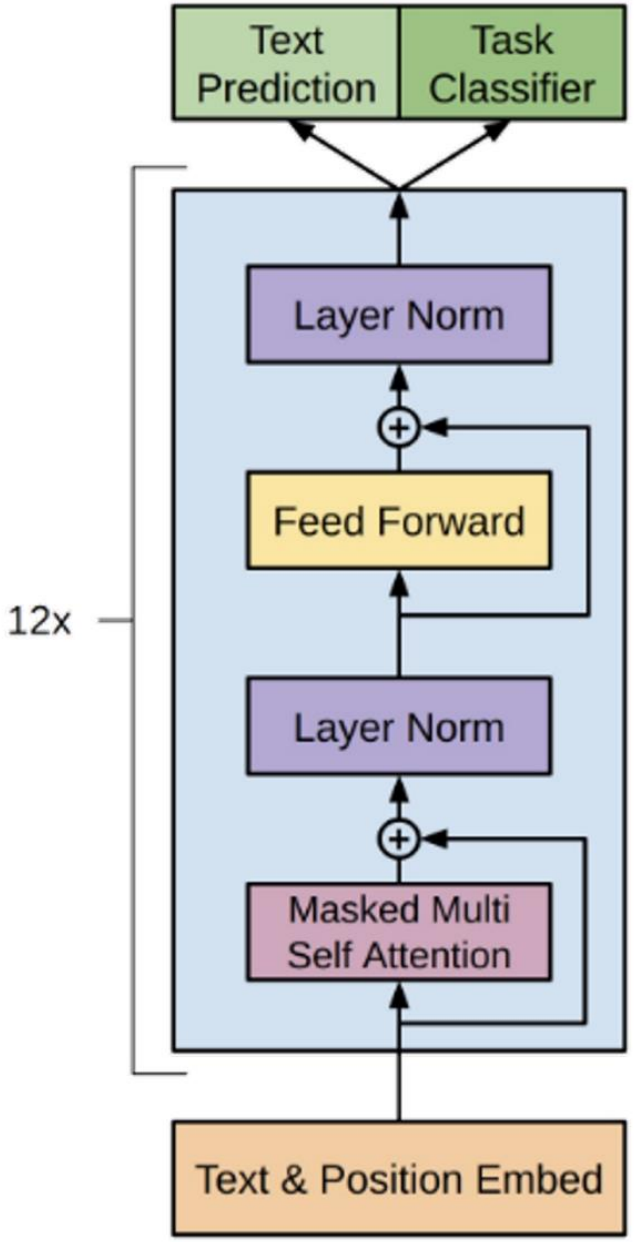
Layer norm(s): (768 * 2) * 2

FF-NN: 1st 768 => 3072, 2nd 3072 => 768

768 * 3072 + 3072 * 768 = (3072 * 768) * 2

= 7,080,960 for one layer => 7,080,960*12

84,971,520



Which of the following is NOT a Parameter-Efficient Fine-Tuning (PEFT) technique?

a) Dynamic quantization
b) Adapter layers
c) Prompt tuning
d) LoRA

Which of the following is NOT a Parameter-Efficient Fine-Tuning (PEFT) technique?

a) Dynamic quantization
b) Adapter layers
c) Prompt tuning
d) LoRA

- The following vectors for the input tokens are given
- Inputs: [2,0], [0,2]
- The following matrices W_q, W_k, W_v are given:

W_q = [[0,1,2],[0,0,2]]
W_k = [[1,2,2],[1,0,1]]
W_v = [[1,2,0],[0,0,1]]

- Compute $\text{Attention}(\text{Q},\text{K},\text{V}) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, for the first token

- T1 [2, 0]
- T2 [0, 2]

Q = T1 x Wq = [0 2 4]

K = [T1, T2] x Wk = [[2,4,4], [2,0,2]]

V = [T1, T2] x Wv = [[2,4,0], [0,0,2]]

Inputs: [2,0], [0,2]

W_q = [[0,1,2], [0,0,2]]

W_k = [[1,2,2], [1,0,1]]

W_v = [[1,2,0], [0,0,1]]

$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- T1 [2, 0]
- T2 [0, 2]

Q = T1 x Wq = [0 2 4]

K = [T1, T2] x Wk = [[2,4,4], [2,0,2]]

V = [T1, T2] x Wv = [[2,4,0], [0,0,2]]

QK^T

2 2

4 0

4 2

0 2 4 **[24, 8] =>** $\frac{[24, 8]}{\sqrt{3}} = \left[\frac{24}{\sqrt{3}}, \frac{8}{\sqrt{3}}\right]$

Inputs: [2,0], [0,2]

W_q = [[0,1,2],
[0,0,2]]

W_k = [[1,2,2],
[1,0,1]]

W_v = [[1,2,0],
[0,0,1]]

$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\frac{QK^T}{\sqrt{d_k}} = \left[\frac{24}{\sqrt{3}}, \frac{8}{\sqrt{3}}\right] = [\,a, b\,]$$

Inputs: [2,0], [0,2]

W_q = [[0,1,2],[0,0,2]]
W_k = [[1,2,2],[1,0,1]]
W_v = [[1,2,0],[0,0,1]]

$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\frac{24}{\sqrt{3}} => \frac{e^x}{\sum ej} = \frac{e^a}{e^a+e^b} \approx 0.999$$

$$\frac{8}{\sqrt{3}} = \frac{e^b}{e^a+e^b} \approx 0.001$$

$$= \left[\frac{e^a}{e^a+e^b}, \frac{e^b}{e^a+e^b}\right] * V \rightarrow \left[2* \frac{e^a}{e^a+e^b}, 4* \frac{e^a}{e^a+e^b}, 0\right] + \left[0, 0, 2* \frac{e^b}{e^a+e^b}\right]$$

$$= \left[2* \frac{e^a}{e^a+e^b}, 4* \frac{e^a}{e^a+e^b}, 2* \frac{e^b}{e^a+e^b}\right]$$

$$= [\,2, 4, 0\,]$$

Which one of these models is a decoder-only?

a) BERT
b) GPT-2
c) T5
d) All options are decoder-only models

Which one of these models is a decoder-only?

a) BERT
b) GPT-2
c) T5
d) All options are decoder-only models

In Causal Language Modeling, the word "causal" means that:

a) The output is generated considering only the current token
b) The output is generated considering only the previous tokens
c) The output is generated considering only the previous and current tokens
d) Only the decoder is used during inference
e) The attention mechanism is adopted
f) The surrounding context must also be used for token generation

In Causal Language Modeling, the word "causal" means that:

a) The output is generated considering only the current token
b) The output is generated considering only the previous tokens
c) The output is generated considering only the previous and current tokens
d) Only the decoder is used during inference
e) The attention mechanism is adopted
f) The surrounding context must also be used for token generation

- A model generates the sentence:
  - "the fast brown fox jumped the bored dog"
- Whereas, the target sentence is:
  - "the quick brown fox jumps over the lazy dog"
- Compute the BLEU-2 score for this generation.
- Remember

$$BLEU-n = BP \cdot \exp\left(\frac{1}{n}\sum_{i}\log(precision_i)\right)$$

$$Brevity\ Penalty = BP = \begin{cases} 1 & if\ g > r \\ e^{\left(1-\frac{r}{g}\right)} & if\ g \leq r \end{cases}$$

- Where $precision_i$ is the precision for the specific i-gram, g and r are the lengths of the generated and reference texts, respectively
- Consider each word to be a token in this case

- “the fast brown fox jumped the bored dog”
- “the quick brown fox jumps over the lazy dog”

1)

G: the, fast, brown, fox, jumped, the, bored, dog

R: the, quick, brown, fox, jumps, over, the, lazy, dog

Precision1 = 5/8

2)

G: (the, fast), (fast, brown), (brown, fox), (fox, jumped), (jumped, the), (the bored), (bored, dog)

R: (the, quick), (quick brown), (brown, fox), (fox, jumps), (jumps, over), (over, the), (the, lazy), (lazy, dog)

Precision2 = 1/7

BLEU-2 = BP * exp(½ * (log(5/8) + log(1/7))

Generated (g) = 8 words

Reference (r) = 9 words

$$BLEU\text{-}n = BP \cdot \exp\left(\frac{1}{n}\sum_i \log(precision_i)\right)$$

$$BP = \begin{cases} 1 & if\ g > r \\ e^{\left(1-\frac{r}{g}\right)} & if\ g \leq r \end{cases}$$

$$BP = e^{\left(1-\frac{r}{g}\right)} = 0.8825$$

$$BLEU - 2 = 0.8825 * \exp\left(\frac{1}{2} * \log(5/8) + \frac{1}{2} * \log(1/7)\right) = 0.2637$$

What is the primary difference between static and dynamic quantization?

a) Dynamic quantization computes scaling values for each activation during inference
b) Dynamic quantization applies to weights only, while static applies to weights and activations
c) Static quantization is slower but less accurate than dynamic quantization
d) Dynamic Quantization requires a calibration step before the quantization
e) None of the others

What is the primary difference between static and dynamic quantization?

a) Dynamic quantization computes scaling values for each activation during inference

b) Dynamic quantization ~~applies to weights only~~, while static applies to weights and activations

c) Static quantization is ~~slower~~ but less accurate than dynamic quantization

d) Dynamic Quantization ~~requires a calibration step~~ before the quantization

e) None of the others

- You are given the following corpus of text: abababab
- Apply BPE to extract 3 tokens (excluding a and b). Write the tokens and their extraction frequency (i.e., the frequency computed at the step when they were merged).
- Note: when doing substitution, work from left to right

Abababab

|a|b|a|b|a|b|a|b|

1) ab: 4 ba: 3
2) (extract ab:4 ➔ X)

Abababab

|a|b|a|b|a|b|a|b|

1) ab: 4 ba: 3
2) (extract ab:4 ➔ X)

|ab|ab|ab|ab|

1) abab: 3
2) (extract abab:3 ➔ Y)

Abababab

|a|b|a|b|a|b|a|b|

1) ab: 4 ba: 3
2) (extract ab:4 ➔ X)

|ab|ab|ab|ab|

1) abab: 3
2) (extract abab:3 ➔ Y)

Abababab

|a|b|a|b|a|b|a|b|

1) ab: 4 ba: 3
2) (extract ab:4 ➔ X)

|ab|ab|ab|ab|

1) abab: 3
2) (extract abab:3 ➔ Y)

|abab|abab|

1) abababab: 1
2) Extract abababab: 1 ➔ Z

In LoRA, what does the process assume about the weight update matrix?

a) That it is low rank or is near low rank
b) That it replaces the initial weight matrix
c) That it is initialized with numbers taken from a normal distribution $\mathcal{N}(0, 1)$
d) That it is fixed (or frozen) throughout training

In LoRA, what does the process assume about the weight update matrix?

a) That it is low rank or is near low rank
b) That it replaces the initial weight matrix
c) That it is initialized with numbers taken from a normal distribution $\mathcal{N}(0, 1)$
d) That it is fixed (or frozen) throughout training

Which one of these is a deterministic sampling approach? Select all correct answers

a) Temperature sampling with T=0

b) Temperature sampling, with $0 < T < 1$

c) Temperature sampling, with $T > 1$

d) Beam Search

e) Greedy sampling

f) Top-k sampling

g) Top-p sampling

Which one of these is a deterministic sampling approach? Select all correct answers

a) Temperature sampling with T=0

b) Temperature sampling, with $0 < T < 1$

c) Temperature sampling, with $T > 1$

d) Beam Search

e) Greedy sampling

f) Top-k sampling

g) Top-p sampling