Karpathy Inference for Llama-2

Ramesh Subramonian

December 27, 2025

Abstract

Step-by-step re-implementation of Karpathy code for llama2.

1 XXX

1.1 Conventions

If T is a table of dimensions $n_1 \times n_2$, then T_i is a vector of length n_2

1.2 Glossary

2 Utilities

2.1 rmsnorm

Input Arguments in Table 2

1. $\alpha = ((\sum_i {x_i}^2)/n) + \epsilon$
2. $o_i \leftarrow \frac{w_i \times x_i}{\alpha}$

Abbreviation	**C code**	XX	**Dimensions**
n_V	`vocab_size`	0	
n_D	`dim`	0	
n_H	`n_heads`	0	
s_H	`head_size`	0	
n_{HKV}	`n_kv_heads`	0	
n_D	`dim`	0	
n'_D	`ispc_dim`	0	
S_{kc}	key_cache		
S_{vc}	value_cache		
W_t	token embedding table	2	$n_V \times n_D$
W_{ra}	rms att weight	2	$n_L \times n_D$
W_q	w_q	3	$n_L \times n_D \times (n_H \times s_H)$
W_k	w_k	3	$n_L \times n_D \times (n_{HK} \times s_H)$
W_v	w_v	3	$n_L \times n_D \times (n_{HK} \times s_H)$
W_o	w_o	3	$n_L \times (n_H \times s_H) \times n_D$

Table 1: Mapping math notation to C code

Argument	**Type**
x	`float` vector of length n
w	`float` vector of length n
n	integer

Table 2: Arguments for rmsnorm
Output arguments

1. o, `float` vector of length n

$x \leftarrow W_t[t]$

for each layer$l$ **do**
$x_b \leftarrow \mathrm{r}msnorm(x, W_{ra}[l]$
$x_{kc} \leftarrow (S_{kc}[l][p]$
$x_{vc} \leftarrow (S_{vc}[l][p]$
$XXX \leftarrow \mathrm{m}atmul(x_{kc}, x_b, W_{wk}[l],)$ SOME JUNK
endfor

Figure 1: Pseudo-code for forward

2.2 softmax

Arguments in Table 3

Argument	Type
x	`float` vector of length n
n	integer

Table 3: Arguments for softmax

1. $m = \mathrm{m}ax_{i=0}^{i=n-1} x_i$
2. $\forall_{i=0}^{i=n-1} x_i \leftarrow e^{x_i - m}$
3. $s = \sum_{i=0}^{i=n-1} x_i$
4. $\forall_{i=0}^{i=n-1} x_i \leftarrow \frac{x_i}{s}$

3 Forward

Arguments in Table 4

Argument	Type	Comments
T	`Transformer`	pointer
t	integer	token $0 \leq t < n_V$
p	integer	pos

Table 4: Arguments for forward