
Karpathy Inference for Llama-2

Ramesh Subramonian

December 20, 2025

Abstract

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

1 Utilities

1.1 rmsnorm

Arguments in Table 1

Argument	Type
o	float vector of length n
x	float vector of length n
w	float vector of length n
n	integer

Table 1: Arguments for rmsnorm

1. $s_1 = \sum_{i=0}^{i=n-1} x_i \times x_i$
2. $s_2 = s_1/n$
3. $s_3 = s_2 + \epsilon$
4. $s = \frac{1}{s_3}$
5. $\forall_{i=0}^{i=n-1} o_i \leftarrow s \times w_i \times x_i$

1.2 softmax

Arguments in Table 2

Argument	Type
x	float vector of length n
n	integer

Table 2: Arguments for softmax

1. $m = \max_{i=0}^{i=n-1} x_i$
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. $\forall_{i=0}^{i=n-1} x_i \leftarrow \frac{x_i}{s}$