Gemma3, Architecture and Mathematical Foundations

Saman Pour

December 16, 2025

Most of the time, we hear about new model releases, new innovations in them or new way of passing around the data; but in reallity very few people know what really happens inside a model.

Everyone starts to refer very general architectural images created years back. The classic can be the widely used Figure 1 from "Attention Is All You Need" [1] paper.

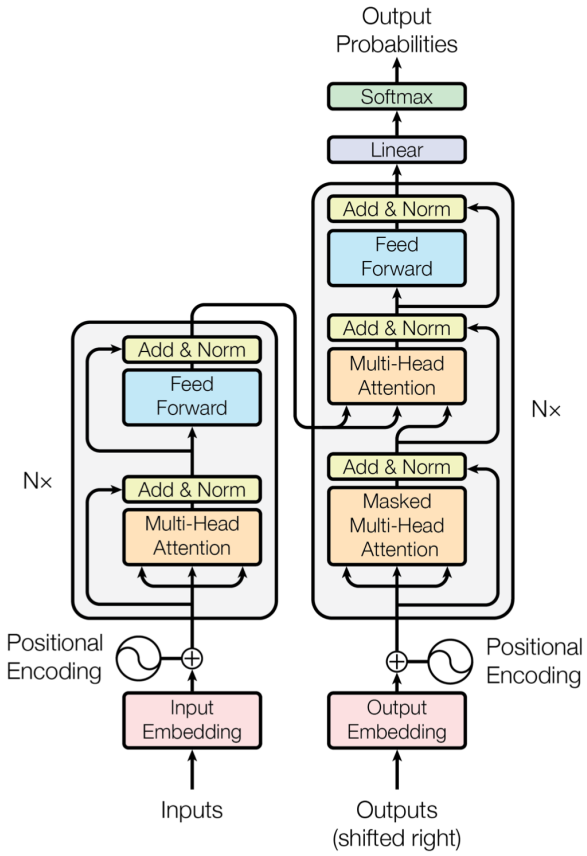

Figure 1: The Transformer - model architecture

Through my work in the domain and studies I've found that the general architecture won't cut it. We need mathematical representation. In addition to that, we need reference implementation that we can use to conduct numerics debugging. A small numerical discripency in the fourth decimal point, can easily turn into catastrophy in output generation when repeated through tens of layers in the decoder.

You might say; hey, just the mathematical description and representation is enought!

I would say the accurate implementation matters too. For example in RMSNorm, Llama does `x.to(float16) * w` whilst Gemma does `(x.to(float) * w).to(float16)`. The former rounds the results, while the later performs the calculation in 32 bit precision first and then rounds it. This slight difference in implementation makes model be better in higher context lenghts.

And that's my motivation for creating metuculusly numerical checked implementation and share it along reports like this full of mathematical background of the subject.

I hope you enjoy this report as much as I enjoyed writing it.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.