Llama

Paper

- https://arxiv.org/pdf/2302.13971.pdf

Pretraining

- datawset
  - English Common Crawl
  - C4
  - Github
  - Wikipedia
  - Gutenberg
  - Arxiv
  - Stack exchange
- Tokenizers this data using BPE algorithm
  - Byte pair encoding algorithm https://arxiv.org/pdf/1508.07909.pdf from sentence piece implementation https://arxiv.org/pdf/1808.06226.pdf

Architecture

- Based on transformer architecture
- Improved on various fields like
  - Pre-normalization
  - SwigLu activation
  - Rotary embeddings

Optimizer

- Used adam

Conclusion

- Nothing very novel besides the training method. The architecture is very standard