ReaLLM ASIC

Make Your Own Lightweight LLMs

Outline of the Tutorial

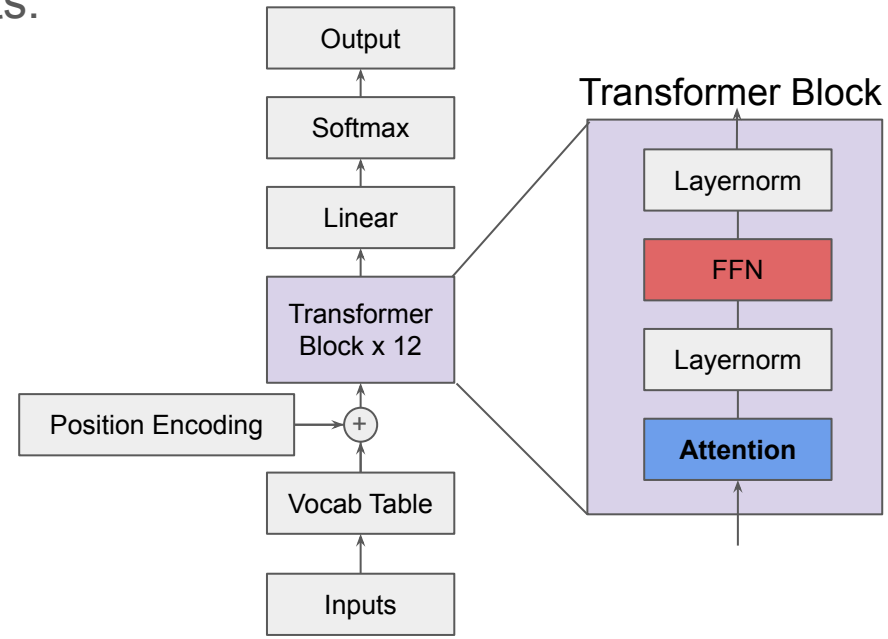
- Intro
  - LLMs and Applications
  - Recap of LLM architecture
- Hands-on AI From Scratch:
  - Building and training a custom lightweight LLM
  - Data preparation and preprocessing
  - Model training options and optimization
- Checkpoints and Finetuning

Applications Limited Only by Datasets

One architecture, limited only by datasets:

- Translation
- Poetry
- Music
- Robotics Motion
- Cooking Recipes
- Editing and Writing Reports

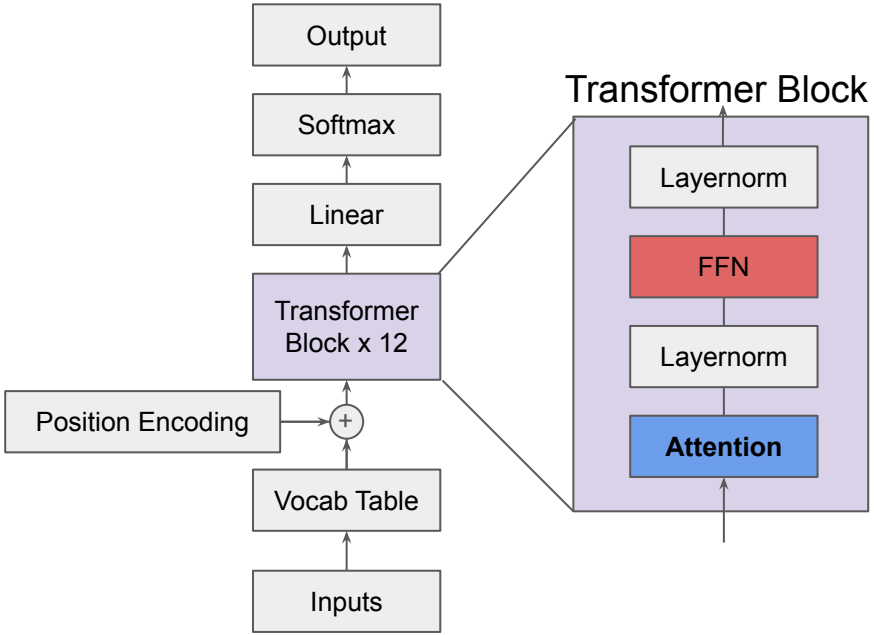
LLM Transformer Architecture



Main Hyperparameters

- “Height” - Number of Layers
  - Deep Networks -> Abstract Knowledge
  - Linearly increases size of network
- “Width” - Dimensions per Token
  - Better Per Token Understanding
  - Non-linear increase in size

LLM Transformer Architecture



Datasets

We are limited by what we can train on.

- Music
  - We'll work with JS Bach
- Language
  - Example with Japanese and English
- Generation
  - Example with Shakespeare

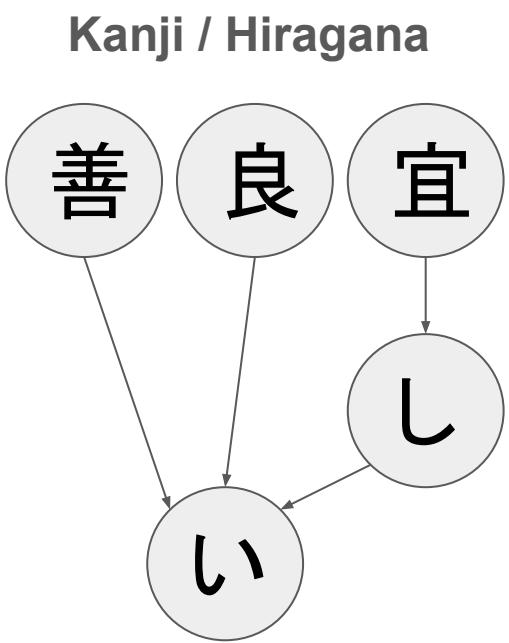
Tokenization Speed-Up Example

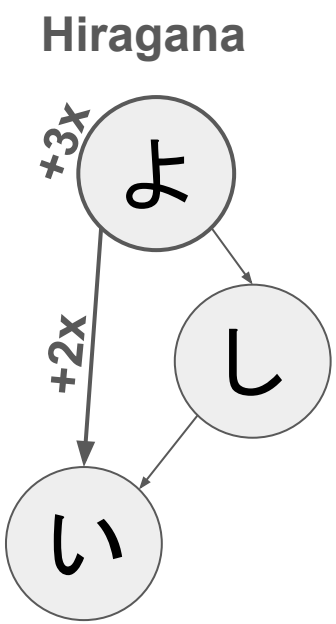
Tokens need to gain “Experience Points”:

- Higher frequency of a token in the dataset then better the LLM will learn about it.

Language Example:

- Preprocessing Kanji -> Hiragana:
  - Faster training (smaller model)
  - Faster accumulation of “experience points”





Colab

Workshop Colab

- Workshop colab
  - based on popular “nanoGPT” framework

- We’ll migrate to the Colab for the remainder of the workshop.