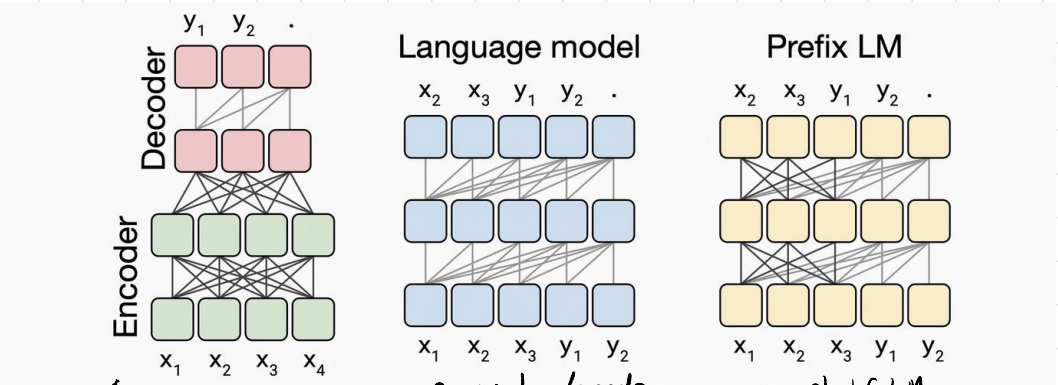


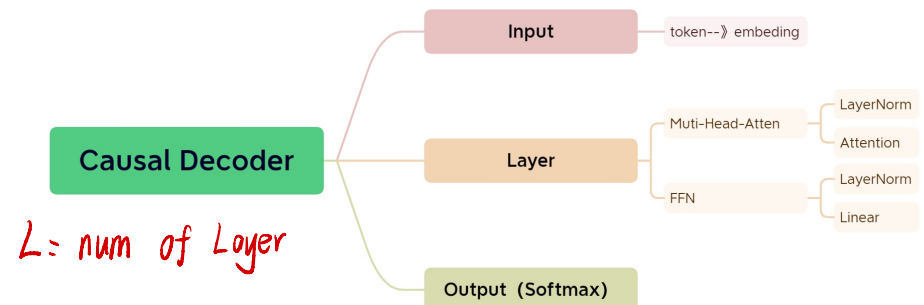
等价于 Transformer

Causal decoder
(Next token prediction)

最常见的架构

LLam系列

ChatGLM
系列



L: num of Layer

H: hidden dim

V: Vocab size

对于 Attention: $\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$

代码中有三个投影 nn.Linear(hidden_dim, hidden_dim)

单个线性层参数为: h^2+h

三个投影+输出Leor $\Rightarrow$ $4h^2+4h$

对于FFN中Linear

维度先升后降 $\Rightarrow [b, s, h] \to [b, s, 4h] \to [b, s, h]$

权重 $\Rightarrow$ $[h, 4h]$ $[4h, h]$

$\Rightarrow$ 计算量 $\Rightarrow$ $4h^2 + 4h + 4h^2 + h \Rightarrow 8h^2 + 5h$

对于LayerNorm $\Rightarrow$ $2h$

对于input与output一般有tie weight：正常为 Vh

总参数为：$(4h^2 + 4h + 8h^2 + 5h + 2h \times 2) \cdot L + Vh$

$\Rightarrow 12Lh^2 + 13Lh + Vh$

一般h比较大，近似为 $12Lh^2$