自然語言處理與應用
Natural Language Processing and Applications

Mixture of Experts (MoE)

Instructor: 林英嘉 (Ying-Jia Lin)
2025/06/01

Course GitHub

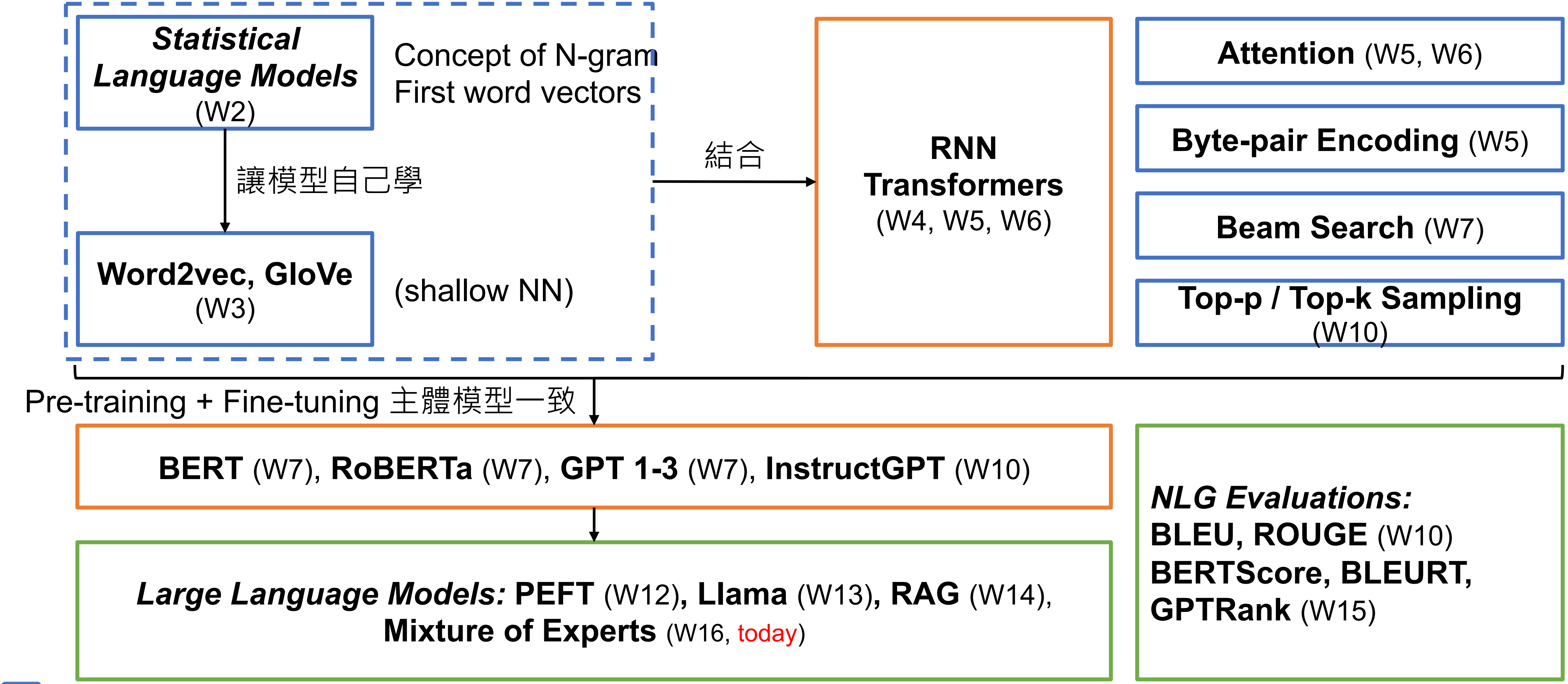
Slido # NLP_0601

Outline

- 學期主題回顧
- Introduction to Mixture of Experts

學期主題回顧

Road Map of Natural Language Processing



Statistical Language Models (W2)

Concept of N-gram
First word vectors

讓模型自己學

Word2vec, GloVe (W3)

(shallow NN)

結合

RNN
Transformers
(W4, W5, W6)

Attention (W5, W6)

Byte-pair Encoding (W5)

Beam Search (W7)

Top-p / Top-k Sampling (W10)

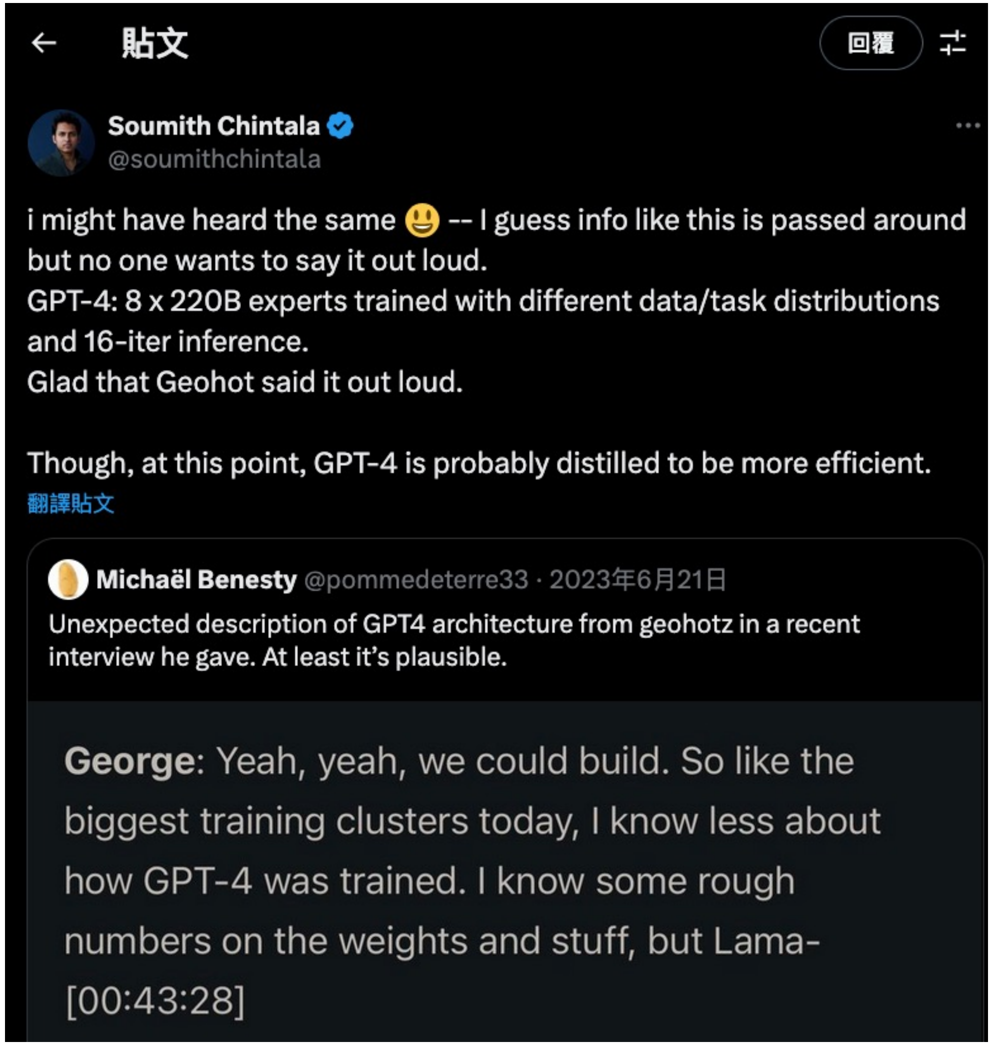
Pre-training + Fine-tuning 主體模型一致

BERT (W7), **RoBERTa** (W7), **GPT 1-3** (W7), **InstructGPT** (W10)

***Large Language Models:* PEFT** (W12)**, Llama** (W13)**, RAG** (W14), **Mixture of Experts** (W16, today)

NLG Evaluations:
BLEU, ROUGE (W10)
BERTScore, BLEURT, GPTRank (W15)

Why do we need to learn MoE?



Soumith Chintala @soumithchintala

i might have heard the same 😃 -- I guess info like this is passed around but no one wants to say it out loud.
GPT-4: 8 x 220B experts trained with different data/task distributions and 16-iter inference.
Glad that Geohot said it out loud.

Though, at this point, GPT-4 is probably distilled to be more efficient.

翻譯貼文

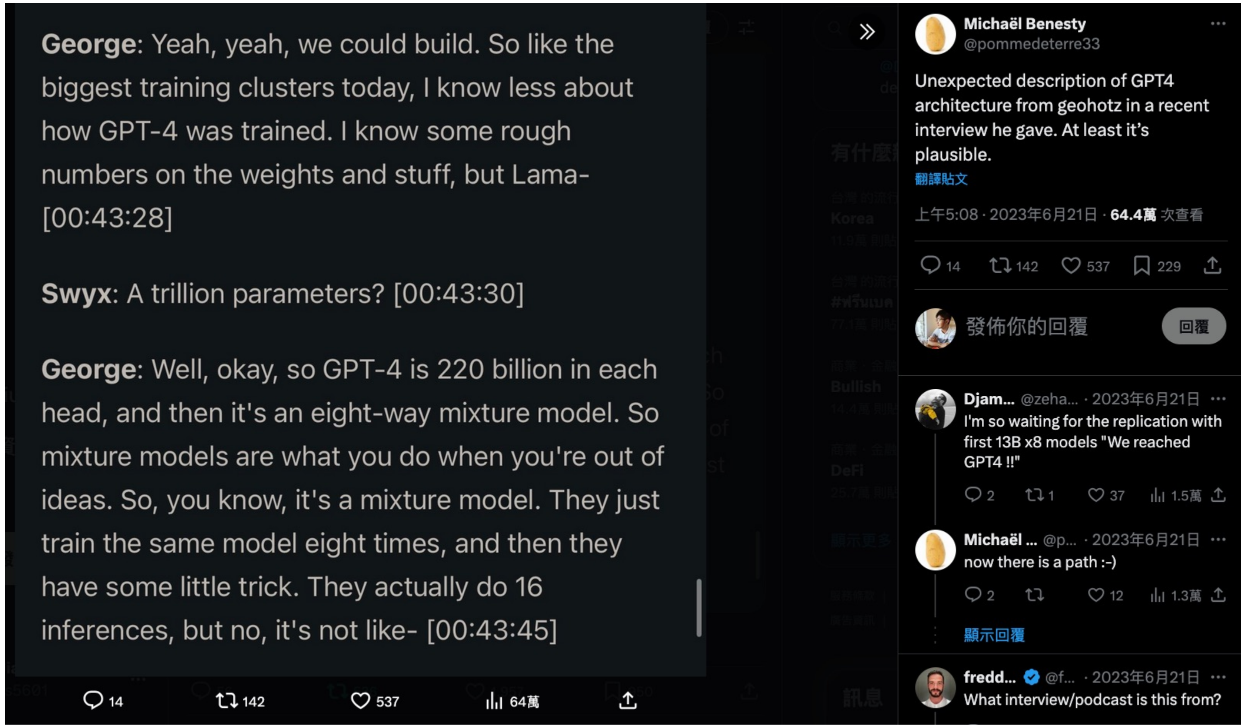
Michaël Benesty @pommedeterre33 · 2023年6月21日

Unexpected description of GPT4 architecture from geohotz in a recent interview he gave. At least it's plausible.

George: Yeah, yeah, we could build. So like the biggest training clusters today, I know less about how GPT-4 was trained. I know some rough numbers on the weights and stuff, but Lama- [00:43:28]

PyTorch 創始者推測 GPT-4 採用 MoE

https://x.com/soumithchintala/status/1671267150101721090



George: Yeah, yeah, we could build. So like the biggest training clusters today, I know less about how GPT-4 was trained. I know some rough numbers on the weights and stuff, but Lama- [00:43:28]

Swyx: A trillion parameters? [00:43:30]

George: Well, okay, so GPT-4 is 220 billion in each head, and then it's an eight-way mixture model. So mixture models are what you do when you're out of ideas. So, you know, it's a mixture model. They just train the same model eight times, and then they have some little trick. They actually do 16 inferences, but no, it's not like- [00:43:45]

Michaël Benesty @pommedeterre33

Unexpected description of GPT4 architecture from geohotz in a recent interview he gave. At least it's plausible.

翻譯貼文

上午5:08 · 2023年6月21日 · **64.4萬** 次查看

Djam... @zeha... · 2023年6月21日

I'm so waiting for the replication with first 13B x8 models "We reached GPT4 !!"

Michaël ... @p... · 2023年6月21日

now there is a path :-)

fredd... @f... · 2023年6月21日

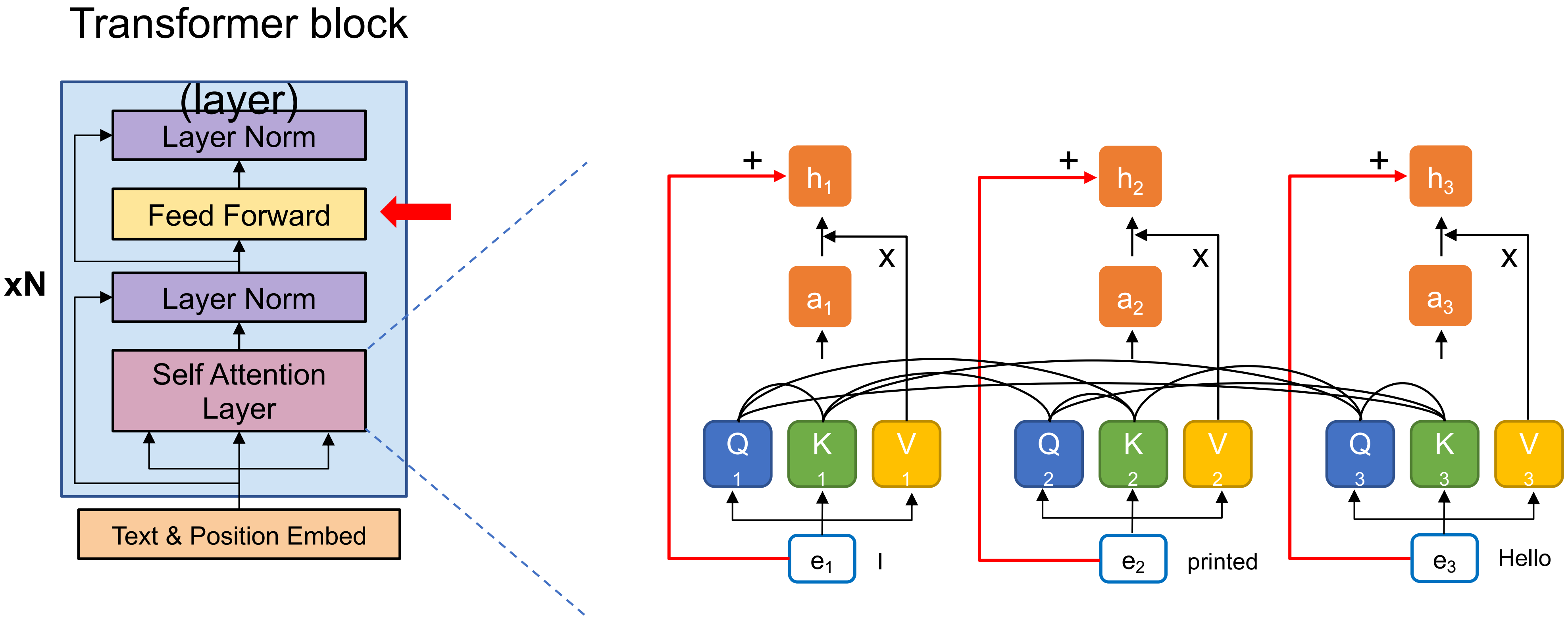
What interview/podcast is this from?

Why do we need to learn MoE?

Model	Release Date	Active Parameters (Total Parameters)	Company
DeepSeek-V3	2024/12/27	37B (671B), 256 experts	DeepSeek
DeepSeek-R1	2025/1/22	37B (671B), 256 experts	DeepSeek
Llama 4 Maverick	2025/4/5	17B (400B), 128 experts	Meta
Mixtral 8x7B	2024/1/8	13B (47B), 8 experts	Mistral AI

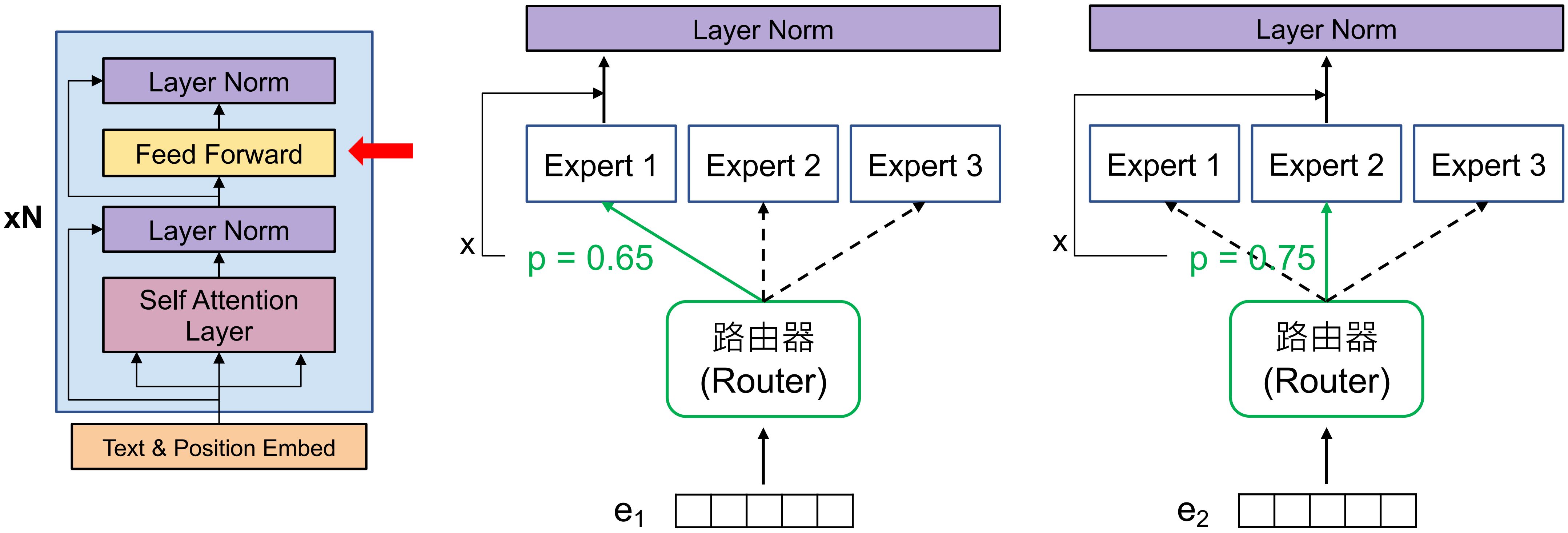
MoE

[Recap] Transformer block



Mixture of Experts (MoE)

Transformer block (layer)

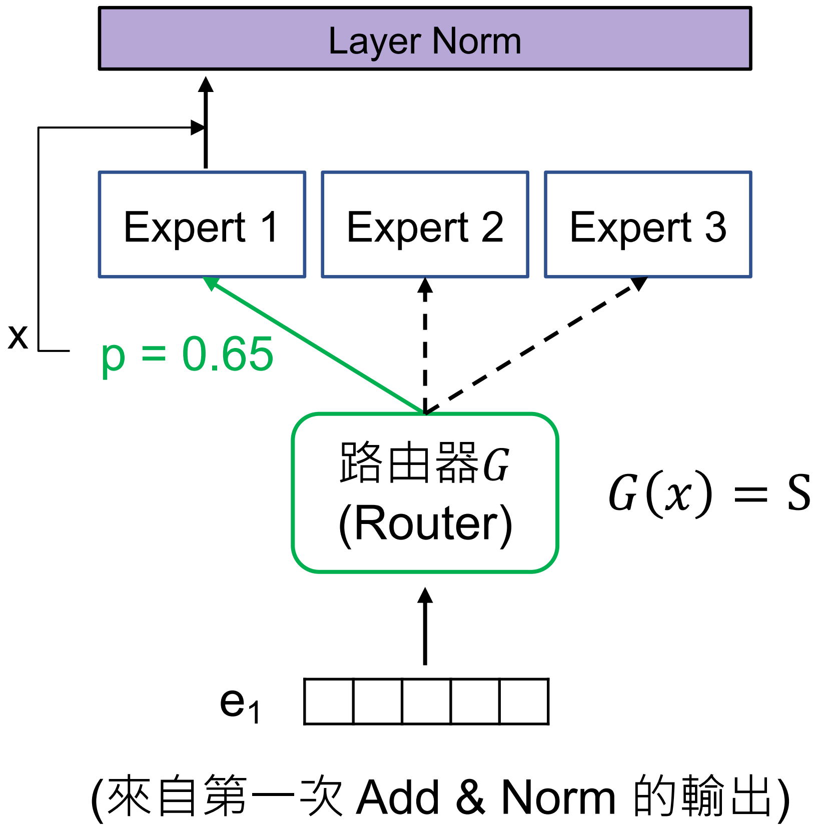


(來自第一次 Add & Norm 的輸出)

(來自第一次 Add & Norm 的輸出)

Router

- Router 是一個小型神經網路，用來決定輸入 token 該被送到哪一個或哪幾個 experts，又稱作 Gating Network



$$G(x) = \mathrm{Softmax}(x \cdot W_g)$$

$W_g \in \mathbb{R}^{d \times E}$

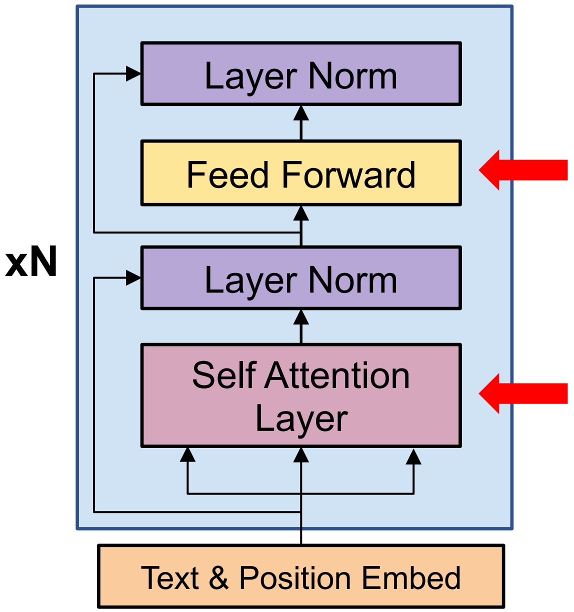
d: x 的向量維度

E: experts 的數量

Expert 是什麼？

- 主要是看你放在哪裡，Expert 就會是一樣的架構

Transformer block (layer)



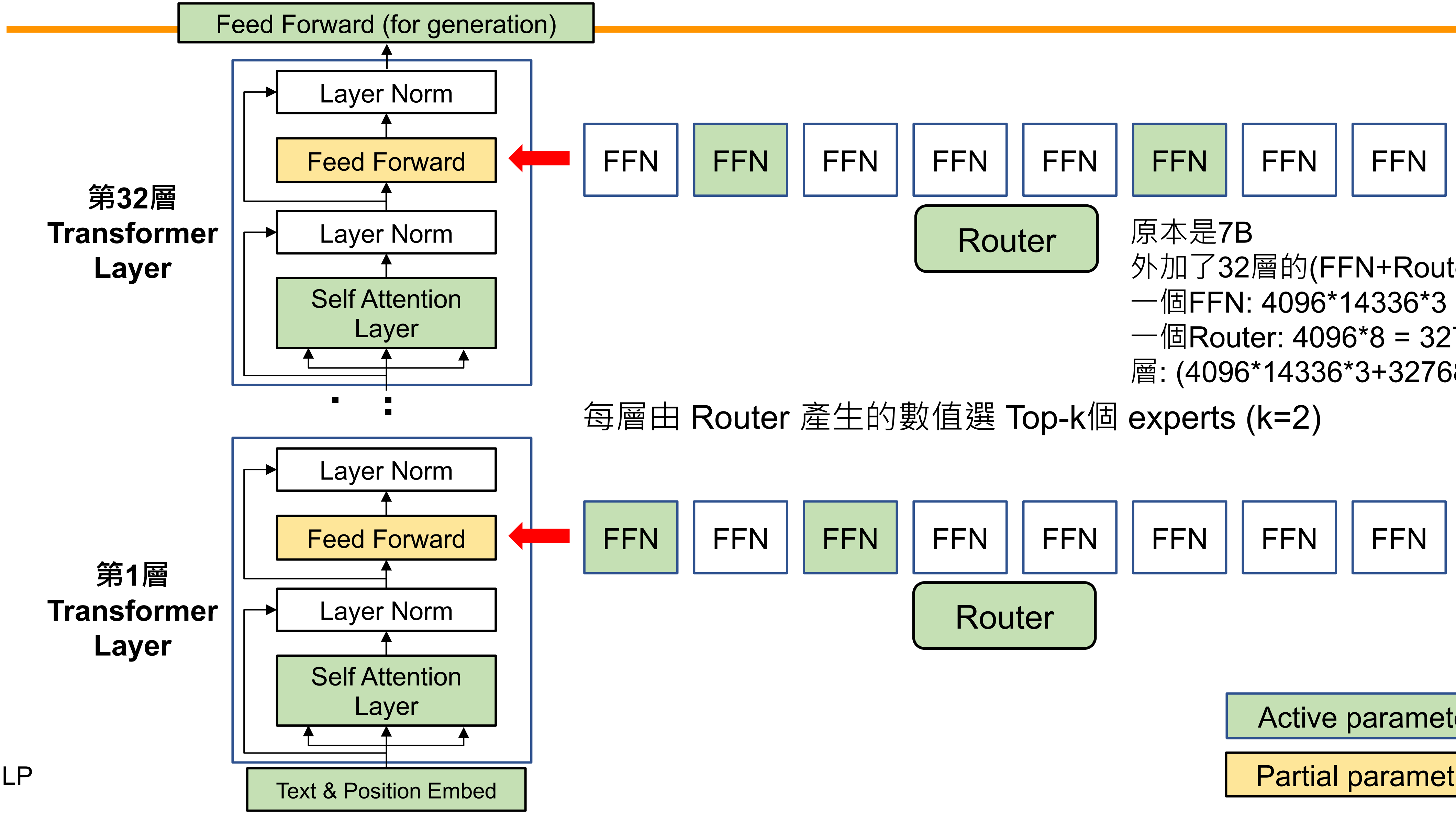
Experts: FFN FFN FFN

E.g., Switch Transformer, Mistral 8x7B, DeepSeek

Experts: Self Attn. Self Attn. Self Attn.

E.g., MoA (Zhang et al., 2022), SwitchHead (Csordás at al., 2024)

Active Parameters (以Mistral 8x7B為例)



原本是7B
外加了32層的(FFN+Router) 參數
一個FFN: 4096*14336*3 (SwiGLU)
一個Router: 4096*8 = 32768
層: (4096*14336*3+32768)*32約5.6B

每層由 Router 產生的數值選 Top-k個 experts (k=2)

🟩 Active parameters

🟨 Partial parameters

Online resources

- Stanford CS25: V4 I Demystifying Mixtral of Experts
- [IBM Technology] What is Mixture of Experts?
- Important papers
  - 近代 MoE 開山之作
    - Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer (Shazeer et al., 2017)
  - MoE on Transformers (T5)
    - Switch Transformers (Fedus et al., 2021)

Thank you!

Instructor: 林英嘉

✉️ yjlin@cgu.edu.tw