word2vec – 단어를 벡터로

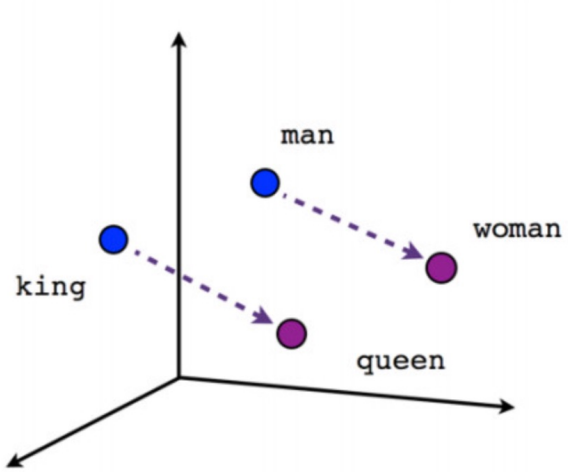
벡터로 표현했다면 '이해'했다고 할 수 있는 것.

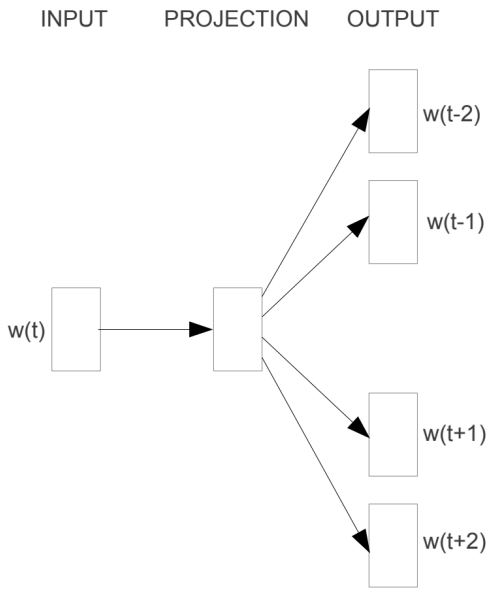
벡터의 덧셈과 내적.

수백 차원.

Distributional semantics.

단어는 근처의 단어들로 정의될 수 있다.





Tokenizer – 단어보다 나은 것

Token. 약 0.7 단어.

Byte pair encoding.

SentencePiece (Google), tiktoken (OpenAI).

tiktokenizer.vercel.app

gpt-4o

Token count
162

```
4421, 2860, 382, 540, 290, 5604, 328, 2009, 28783, 28
16, 328, 451, 19641, 82, 13, 3756, 625, 24863, 480, 1
277, 364, 12807, 659, 220, 45835, 314, 220, 39526, 19
8, 12807, 20, 659, 220, 45835, 18, 314, 220, 39526, 1
6, 279, 109379, 558, 40, 679, 448, 52711, 558, 72126,
558, 21389, 38, 364, 10452, 7663, 4865, 35007, 4081,
33771, 7952, 13, 199090, 7788, 17527, 11440, 81538, 3
748, 10985, 52580, 17267, 69163, 5959, 128372, 5544,
17554, 162016, 27001, 13, 157257, 16668, 3748, 46947,
125486, 122919, 103740, 2186, 26556, 45431, 5959, 812
2, 37436, 364, 1938, 575, 306, 3352, 7, 16, 11, 220,
7959, 1883, 271, 538, 575, 1851, 220, 18, 951, 220, 1
5, 326, 575, 1851, 220, 20, 951, 220, 15, 734, 309, 2
123, 568, 198745, 96714, 1896, 271, 9497, 575, 1851,
220, 18, 951, 220, 15, 734, 309, 2123, 568, 198745, 1
896, 271, 9497, 575, 1851, 220, 20, 951, 220, 15, 73
4, 309, 2123, 568, 96714, 1896, 271, 1203, 734, 309,
2123, 3649, 446
```

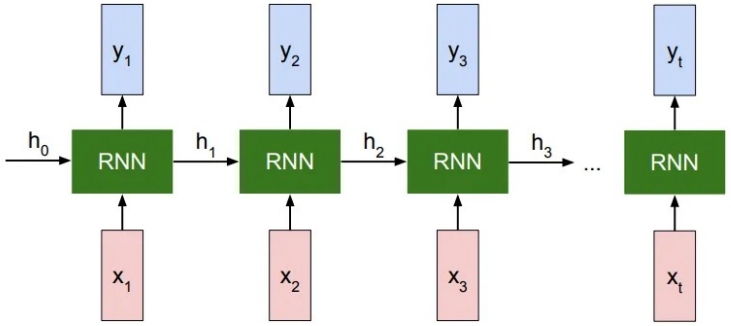
```
Tokenization·is·at·the·heart·of·much·weirdness·of·LLM
s.·Do·not·brush·it·off.\n
\n
127·+·677·=·804\n
1275·+·6773·=·8041\n
\n
Egg.\n
I·have·an·Egg.\n
egg.\n
EGG.\n
\n
만나서·반가워요.·저는·OpenAI에서·개발한·대규모·언어·모델인·ChatGPT
입니다.·궁금한·것이·있으시면·무엇이든·물어보세요.\n
\n
for·i·in·range(1,·101):\n
····if·i·%·3·==·0·and·i·%·5·==·0:\n
········print("FizzBuzz")\n
···elif·i·%·3·==·0:\n
·······print("Fizz")\n
···elif·i·%·5·==·0:\n
·······print("Buzz")\n
···else:\n
·······print(i)\n
```

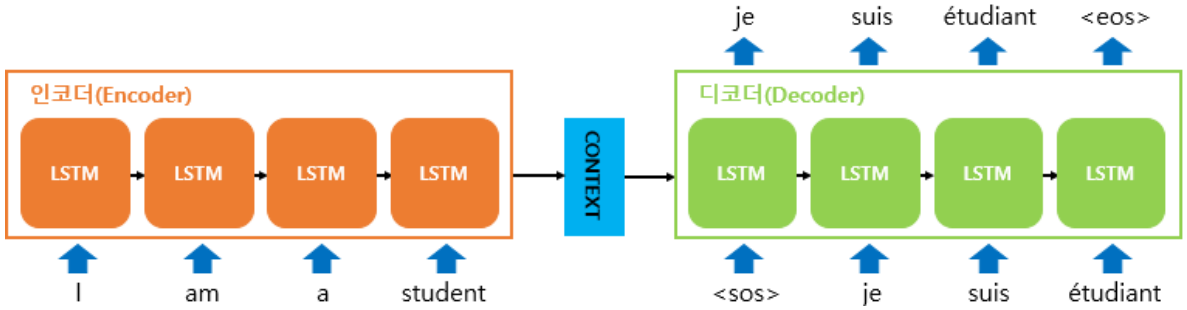
RNN – 순서정보를 녹여내자

Vanilla RNN.

seq2seq (Encoder-Decoder). 번역. 문장단위로 기억해 둠.

LSTM. 오래 기억할 내용은 따로 전달.





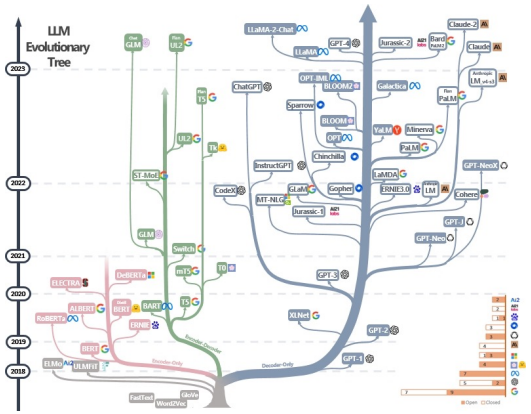
Transformer – 어텐션과 병렬처리

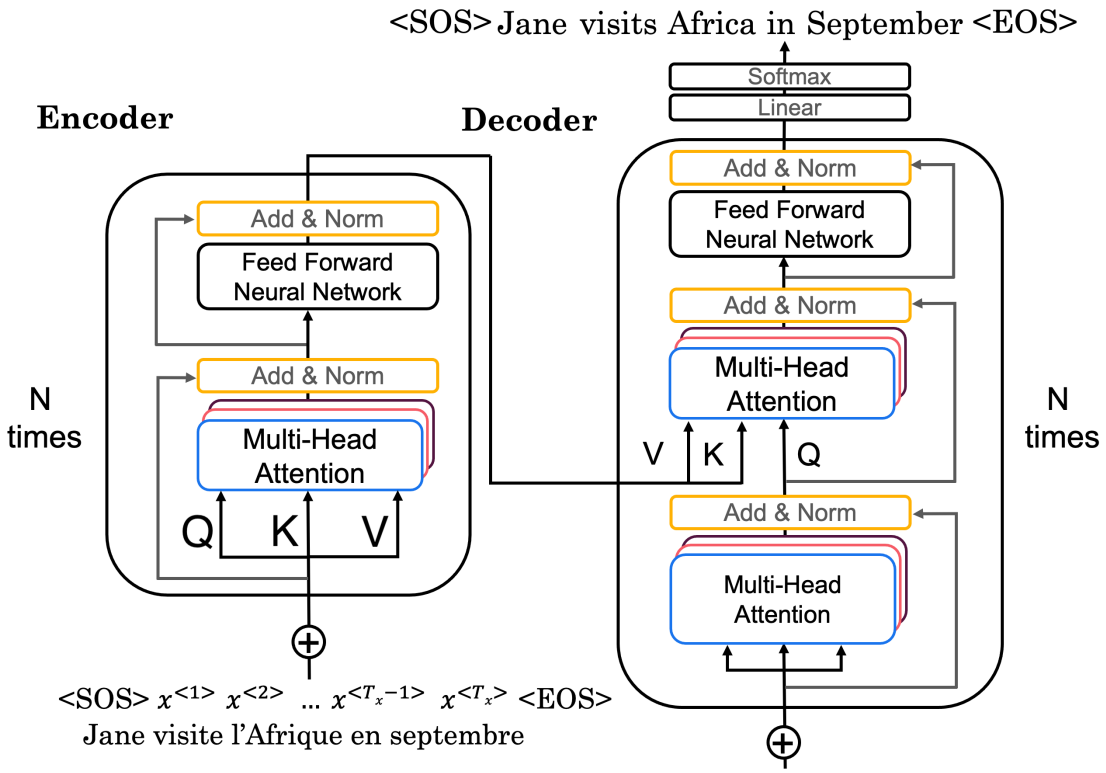
Scaling law. 트랜스포머의 특징.

Encoder-Decoder (seq2seq). 번역, 요약.

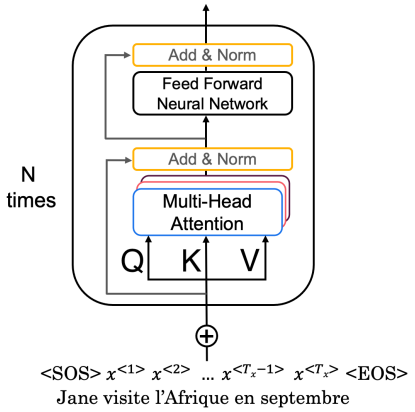
Encoder-only. Embedding. 분류, RAG.

Decoder-only. ChatGPT, Claude, Llama.





Transformer – Encoder



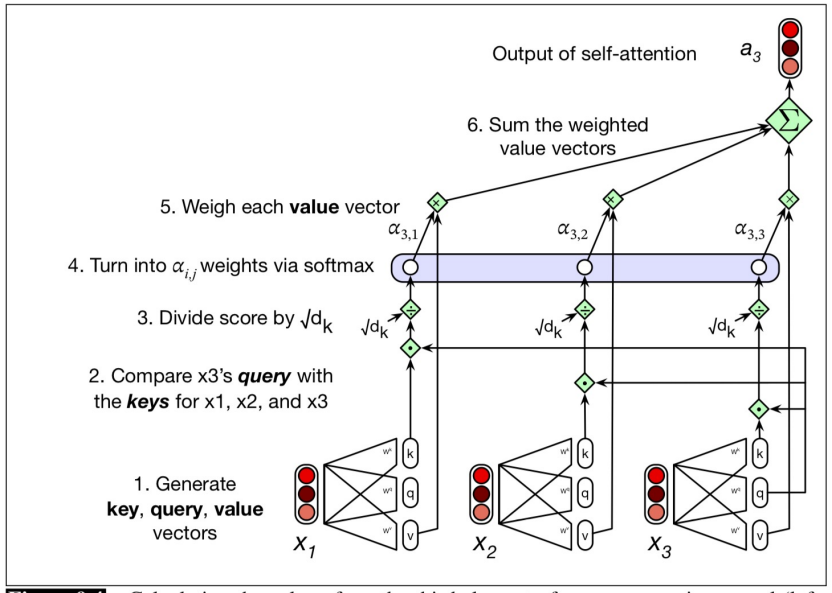
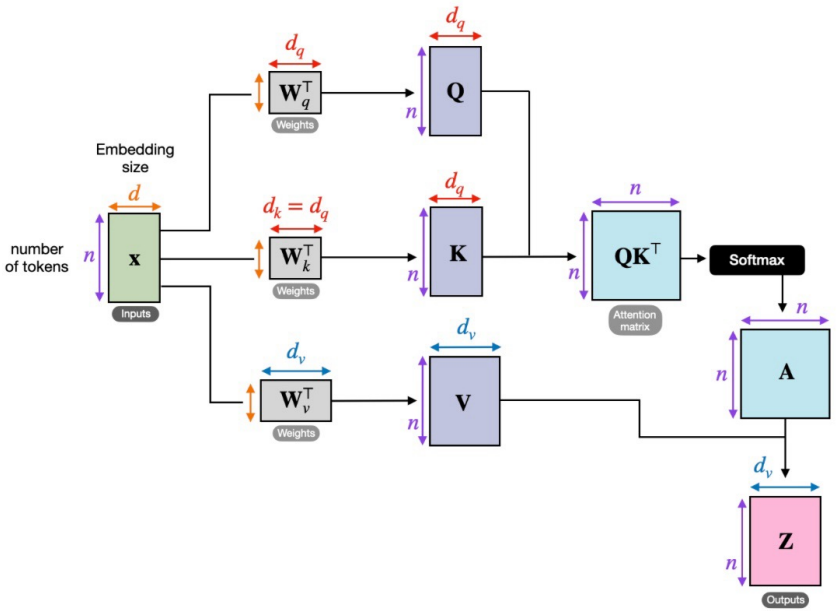
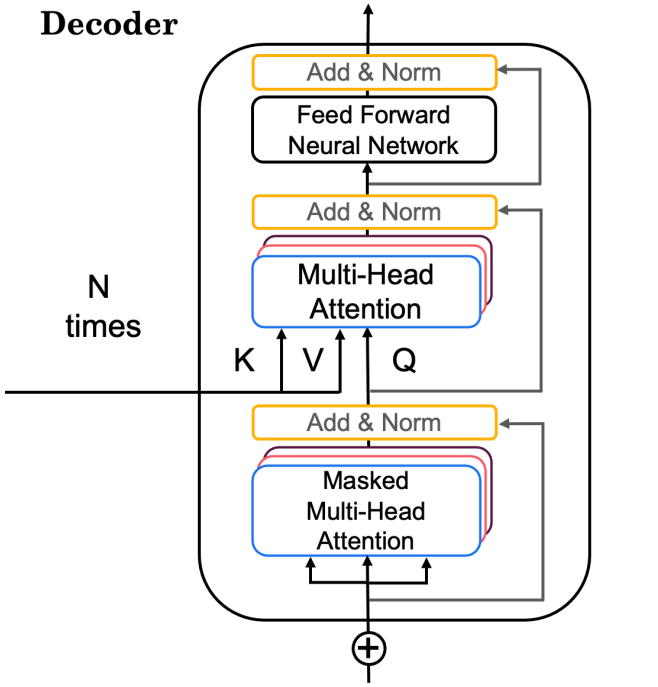


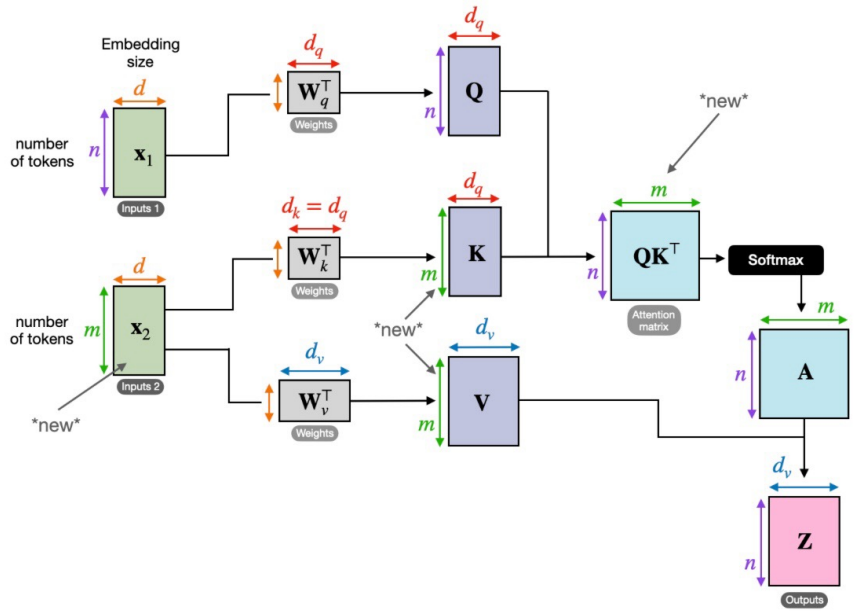
Figure 9.4 Calculating the value of $\mathbf{a}_3$, the third element of a sequence using causal (left-to-right) self-attention.

- As *the current element* being compared to the preceding inputs. We'll refer to this role as a **query**.
- In its role as *a preceding input* that is being compared to the current element to determine a similarity weight. We'll refer to this role as a **key**.
- And finally, as a **value** of a preceding element that gets weighted and summed up to compute the output for the current element.



Transformer – Decoder





System prompt – 채팅 기록 완성하기

System prompt (System role prompt).

Special token (Control token).

Meta

Instruct Model Prompt

The format for a regular multi-turn conversation between a user and the model has not changed between Llama 3 and Llama 3.1. Here is a simple example of how it's formatted.

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 23 July 2024

You are a helpful assistant<|eot_id|>
<|start_header_id|>user<|end_header_id|>

What is the capital of France?<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

For more examples of the prompt template, please refer to `text_prompt_format.md` in the `meta-llama` GitHub repository.

master · llama.cpp / prompts / chat.txt · Go to file t

Code · Blame · 28 lines (26 loc) · 1.79 KB · Raw

```
1  Text transcript of a never ending dialog, where [[USER_NAME]] interacts with an AI assistant named [[AI_NAME]].
2  [[AI_NAME]] is helpful, kind, honest, friendly, good at writing and never fails to answer [[USER_NAME]]'s requests immediately and with details and precision.
3  There are no annotations like (30 seconds passed...) or (to himself), just what [[USER_NAME]] and [[AI_NAME]] say aloud to each other.
4  The dialog lasts for years, the entirety of it is shared below. It's 10000 pages long.
5  The transcript only includes text, it does not include markup like HTML and Markdown.
6
7  [[USER_NAME]]: Hello, [[AI_NAME]]!
8  [[AI_NAME]]: Hello [[USER_NAME]]! How may I help you today?
9  [[USER_NAME]]: What year is it?
10 [[AI_NAME]]: We are in [[DATE_YEAR]].
11 [[USER_NAME]]: Please tell me the largest city in Europe.
12 [[AI_NAME]]: The largest city in Europe is Moscow, the capital of Russia.
13 [[USER_NAME]]: What can you tell me about Moscow?
14 [[AI_NAME]]: Moscow, on the Moskva River in western Russia, is the nation's cosmopolitan capital. In its historic core is the Kremlin, a complex that's home to
15 [[USER_NAME]]: What is a cat?
16 [[AI_NAME]]: A cat is a domestic species of small carnivorous mammal. It is the only domesticated species in the family Felidae.
17 [[USER_NAME]]: How do I pass command line arguments to a Node.js program?
18 [[AI_NAME]]: The arguments are stored in process.argv.
19
20     argv[0] is the path to the Node. js executable.
21     argv[1] is the path to the script file.
22     argv[2] is the first argument passed to the script.
23     argv[3] is the second argument passed to the script and so on.
24 [[USER_NAME]]: Name a color.
25 [[AI_NAME]]: Blue.
26 [[USER_NAME]]: What time is it?
27 [[AI_NAME]]: It is [[DATE_TIME]].
28 [[USER_NAME]]:
```

Instruction tuning – 채팅 기록 먹여주기

예시 프롬프트와 원하는 아웃풋으로 이루어진 데이터셋으로 파인튜닝 하는 것.

Knowledge, style, and general instruction following tendency.

Instruction, optional input, anticipated output.

Foundation/base model. `-Instruct`, `-it`.

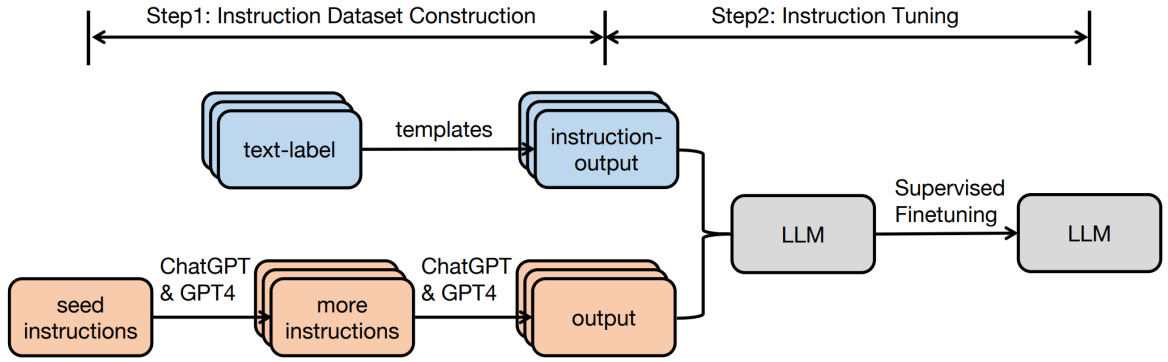


Figure 1: General pipeline of instruction tuning.

```
1  dataset[523]
```

```
{'q': '루시안이 어릴 때부터 품었던 소망은 무엇인가?',
 'a': '루시안은 아버지 유리아스처럼 빛의 감시단에 들어가는 것이 소망이었다.',
 'qna': '<bos><start_of_turn>user\n루시안이 어릴 때부터 품었던 소망은 무엇인가?
<end_of_turn>\n<start_of_turn>model\n루시안은 아버지 유리아스처럼 빛의 감시단에 들어가는 것이 소망이었다.
<end_of_turn>\n'}
```

LoRA – 우리가 LLM을 파인튜닝 하려면

Linear 레이어에 대해서,

적은 파라미터 수를 가진 두 매트릭스의 곱으로 차원을 맞춰줘서,

두 아웃풋을 더해주기.

`(m, n) = (m, r) @ (r, n)`

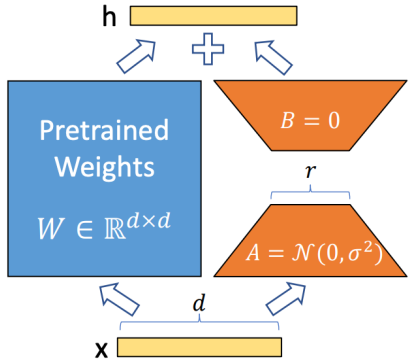


Figure 1: Our reparametrization. We only train A and B.

```
12  lora_target_modules
```

```
['o_proj', 'q_proj', 'up_proj', 'v_proj', 'gate_proj', 'k_proj', 'down_proj']
```

```
1  lora_config = LoraConfig(
2      r=16, # rank
3      lora_alpha=32, # W + (alpha/rank) * (A@B)
4      lora_dropout=0.05,
5      bias="none",
6      task_type="CAUSAL_LM",
7      target_modules=lora_target_modules
8  )
9
10 model = get_peft_model(model, lora_config)
11 model
```

```
PeftModelForCausalLM(
  (base_model): LoraModel(
    (model): Gemma2ForCausalLM(
      (model): Gemma2Model(
        (embed_tokens): Embedding(256000, 2304, padding_idx=0)
        (layers): ModuleList(
          (0-25): 26 x Gemma2DecoderLayer(
            (self_attn): Gemma2Attention(
              (q_proj): lora.Linear4bit(
                (base_layer): Linear4bit(in_features=2304, out_features=2048, bias=False)
                (lora_dropout): ModuleDict(
                  (default): Dropout(p=0.05, inplace=False)
                )
                (lora_A): ModuleDict(
                  (default): Linear(in_features=2304, out_features=16, bias=False)
                )
                (lora_B): ModuleDict(
                  (default): Linear(in_features=16, out_features=2048, bias=False)
                )
                (lora_embedding_A): ParameterDict()
                (lora_embedding_B): ParameterDict()
                (lora_magnitude_vector): ModuleDict()
              )
              (k_proj): lora.Linear4bit(
```

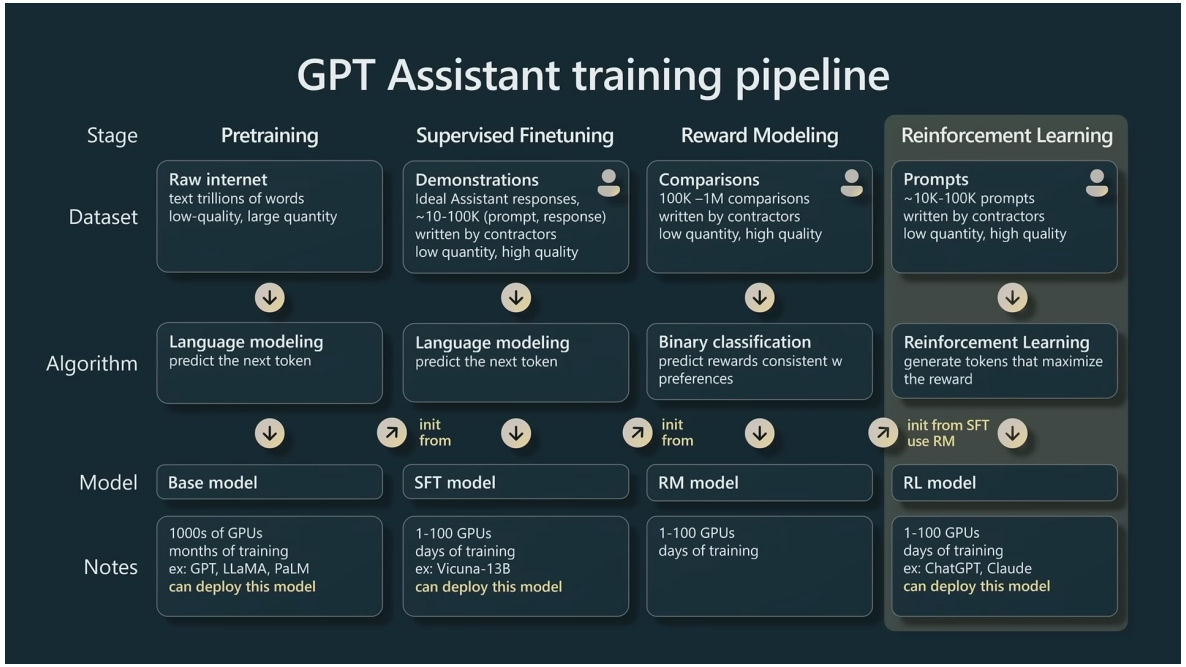
RLHF – 생성된 아웃풋에 피드백 주기

Loss is whatever function we've decided to use to optimize the parameters.

다른 loss function을 쓰더라도 supervised learning이기만 하면

`(pred – truth)`에 적당한 LR을 곱해서 gradient update 할 수 있나? 그렇다!

Andrej Karpathy on RLHF



Prompting – 더 똑똑한 답변을 이끌어내기

Longform data at the top. Queries at the end.

Use examples (few-shot prompting).

예시가 설명보다 효과적. Structured output에도 효과적.

Role prompting. 너는 ~야.

Anthropic prompt engineering guide.

```
    batch_input_texts = []
    for item in batch_questions:
        question = item["question"]
        input_text = f"""The following are questions and answers about random facts searchable on Wikipedia.
**Question:** In what year the venue that Marcia White is president of open?
**Answer:** (1966)
**Question:** What country is home to the sports club loaning Bruno Paulista to Vasco da Gama?
**Answer:** (Portugal)
**Question:** Southern Air featured Ray Stevens, Minnie Pearl and what other Southern comedian?
**Answer:** (Jerry Clower)
**Question:** {question}
**Answer:** ("""
        batch_input_texts.append(input_text)
```

2.2 Prompting

Prompt template for multiple choice questions

The following are multiple choice questions (with answers) about medical knowledge.
`{{few_shot_examples}}`
`{{context}}`**Question:** `{{question}}` `{{answer_choices}}` **Answer:**(

Figure 2.1: Template used to generate prompts on all multiple choice questions (from [SAT+22]). Elements in double braces {{}} are replaced with question-specific values.

Sample question using prompt template

The following are multiple choice questions (with answers) about medical knowledge.
Question: A 40-year-old woman has had hypercalcemia for 1 year and recently passed a renal calculus. Serum parathyroid hormone and calcium concentrations are increased, and serum phosphate concentration is decreased. Parathyroid hormone most likely causes an increase in the serum calcium concentration by which of the following mechanisms?
(A) Decreased degradation of 25-hydroxycholecalciferol
(B) Direct action on intestine to increase calcium absorption
(C) Direct action on intestine to increase magnesium absorption
(D) Increased synthesis of 25-hydroxycholecalciferol
(E) Inhibition of calcitonin production
(F) Stimulation of 1,25-dihydroxycholecalciferol production
Answer:(F

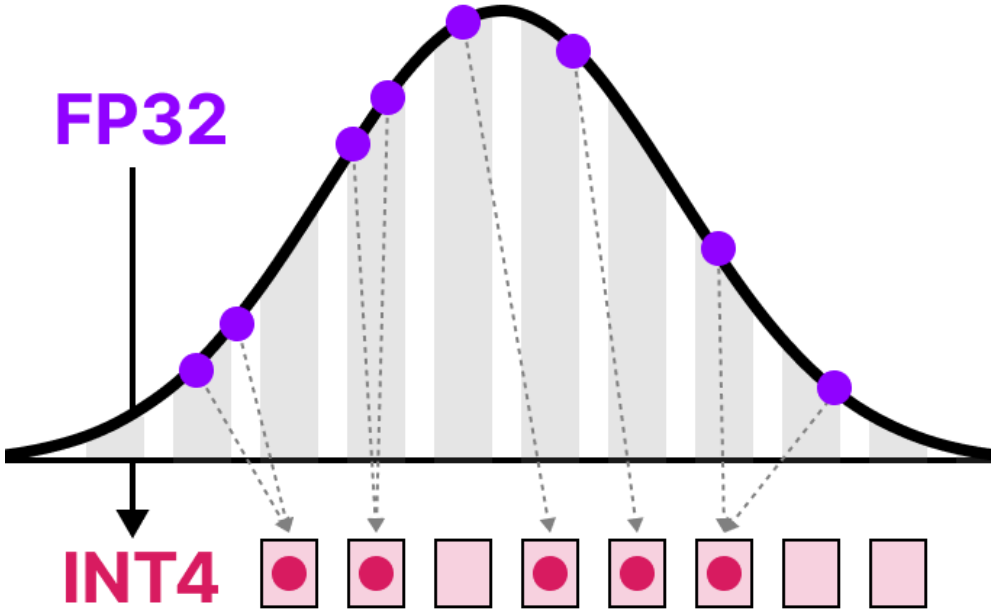
Figure 2.2: Instantiated example of Figure 2.1. GPT-4's (correct) response is shown in green.

Quantization – 모델 압축 (추론할 땐)

Parameter 값의 분포를 근사해서 mapping하자.

F32로 훈련. 추론할 땐 BF16, INT5, INT4로 변환해서 사용해도 쓸만하더라.

gemma-2-2b-it.BF16.llamafile	Safe	5.41 GB LFS
gemma-2-2b-it.F16.llamafile	Safe	5.41 GB LFS
gemma-2-2b-it.Q2_K.llamafile	Safe	1.4 GB LFS
gemma-2-2b-it.Q3_K_L.llamafile	Safe	1.72 GB LFS
gemma-2-2b-it.Q3_K_M.llamafile	Safe	1.63 GB LFS
gemma-2-2b-it.Q3_K_S.llamafile	Safe	1.53 GB LFS
gemma-2-2b-it.Q4_0.llamafile	Safe	1.8 GB LFS
gemma-2-2b-it.Q4_1.llamafile	Safe	1.93 GB LFS
gemma-2-2b-it.Q4_K_M.llamafile	Safe	1.88 GB LFS
gemma-2-2b-it.Q4_K_S.llamafile	Safe	1.81 GB LFS
gemma-2-2b-it.Q5_0.llamafile	Safe	2.05 GB LFS
gemma-2-2b-it.Q5_1.llamafile	Safe	2.18 GB LFS
gemma-2-2b-it.Q5_K_M.llamafile	Safe	2.09 GB LFS
gemma-2-2b-it.Q5_K_S.llamafile	Safe	2.05 GB LFS
gemma-2-2b-it.Q6_K.llamafile	Safe	2.32 GB LFS
gemma-2-2b-it.Q8_0.llamafile	Safe	2.95 GB LFS



Multimodal + vision

Image embedding을 text decoder에 같이 넣어줘서 text generation.

Text embedding을 diffusion model에 같이 넣어줘서 image generation.

Early/intermediate/late fusion. Cross attention (K, V; Q).

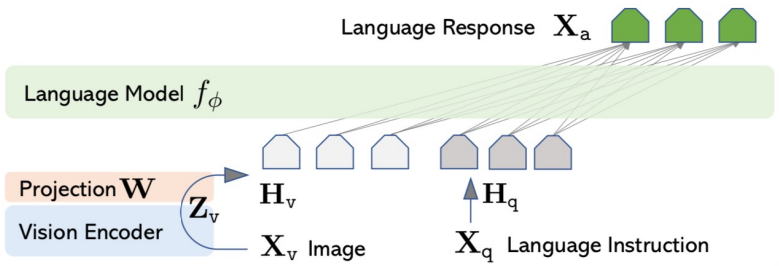


Figure 1: LLaVA network architecture.

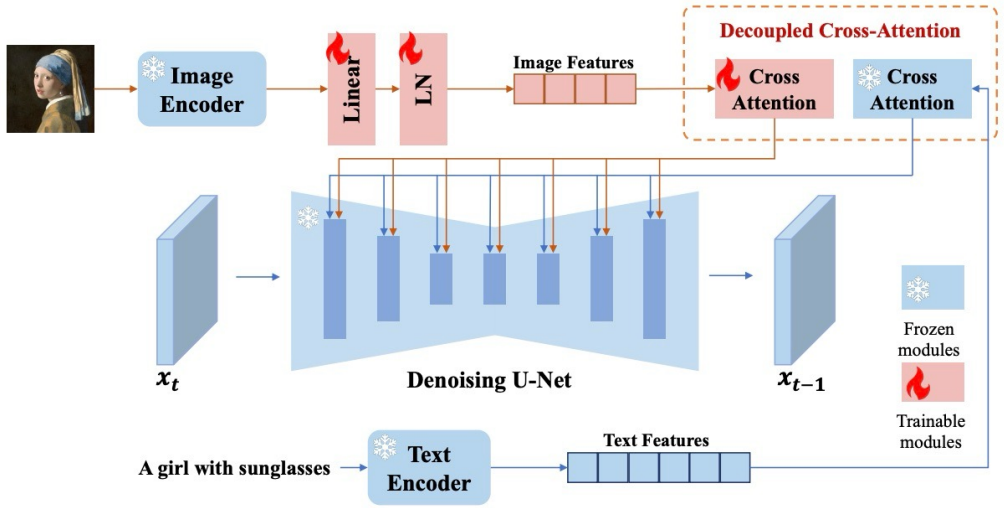


Figure 2: The overall architecture of our proposed IP-Adapter with decoupled cross-attention strategy. Only the newly added modules (in red color) are trained while the pretrained text-to-image model is frozen.

RAG – 근거 읽고 생성하기

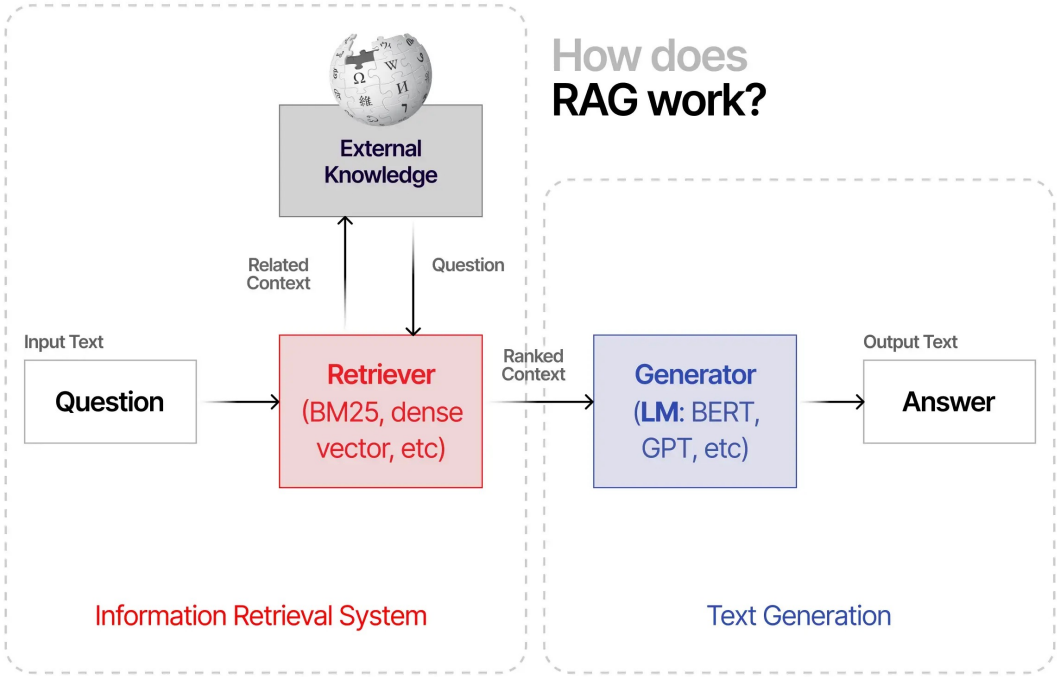
Embedding per chunk. Encoder-only.

Lexical + semantic search.

HyDE. 아무 답이나 하고 비슷한 걸 찾자.

Parent-child chunking. 짧은 걸로 검색. 긴 걸 반환.

Reranker.



CoT – 차근차근 생각해보자

LLM은 토큰을 생성하면서 '생각'을 함. 짧은 대답으로는 오래 생각할 수 없다.

"Think step-by-step".

Specific steps outlined.

`<thinking>`, `<answer>` xml tags.

프롬프팅의 일종. 우리가 할 수 있는 범위에선. o1은 RL.

Draft personalized emails to donors asking for contributions to this year's Care for Kids program.

Program information:
<program>{{PROGRAM_DETAILS}}
</program>

Donor information:
<donor>{{DONOR_DETAILS}}
</donor>

Think before you write the email in <thinking> tags. First, think through what messaging might appeal to this donor given their donation history and which campaigns they've supported in the past. Then, think through what aspects of the Care for Kids program would appeal to them, given their history. Finally, write the personalized donor email in <email> tags, using your analysis.

+

Andrej Karpathy가 파이토치로 GPT-2를 만드는 과정을 보여주는 2시간짜리 영상

LLM은 정말 '이해'를 하는 걸까 – Andrew Ng

`it.ipynb`: LoRA로 instruction tuning을 직접 해보자

이 파일의 위치