A Survey on Multimodal Large Language Models for Anomaly Detection

김규형

CONTENTS

01

Introduction

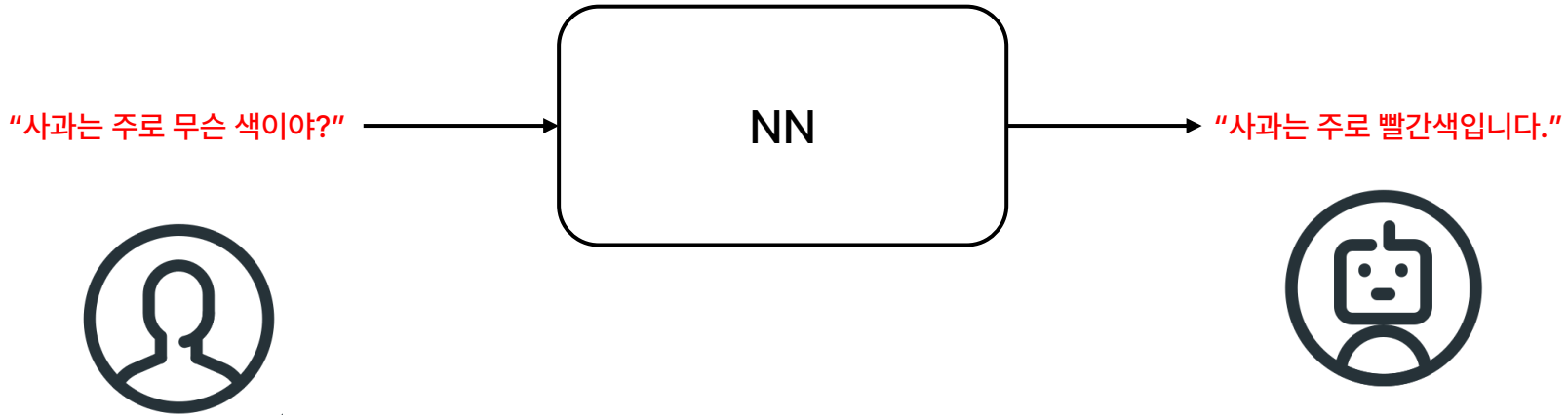
LLM

Language Model (LM)

- 텍스트라는 매개체를 통해 사람과 상호작용 할 수 있는 인공지능 모델

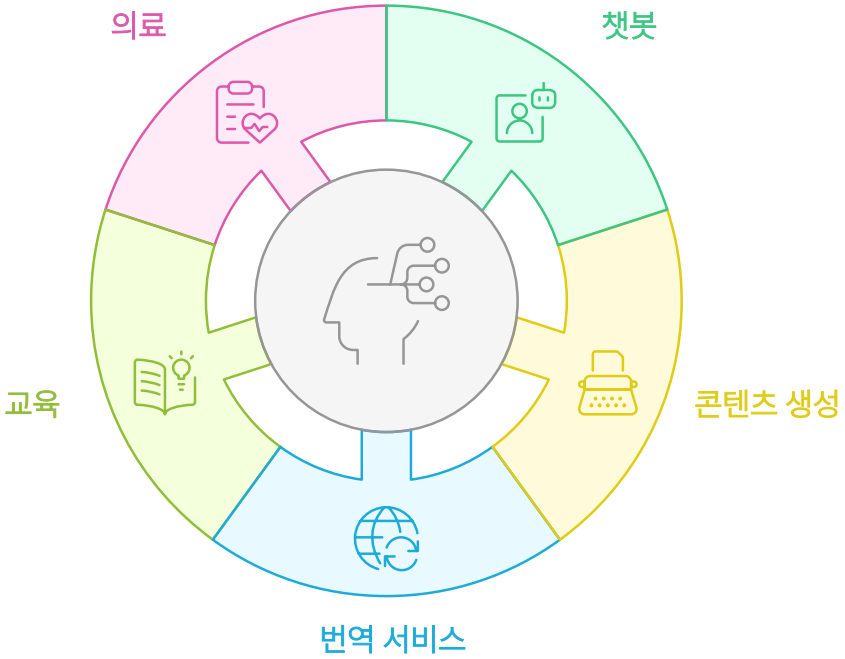
Auto-regressive

- T 시점까지 주어진 단어들을 기반으로 다음 T+1 단어를 예측하도록 학습



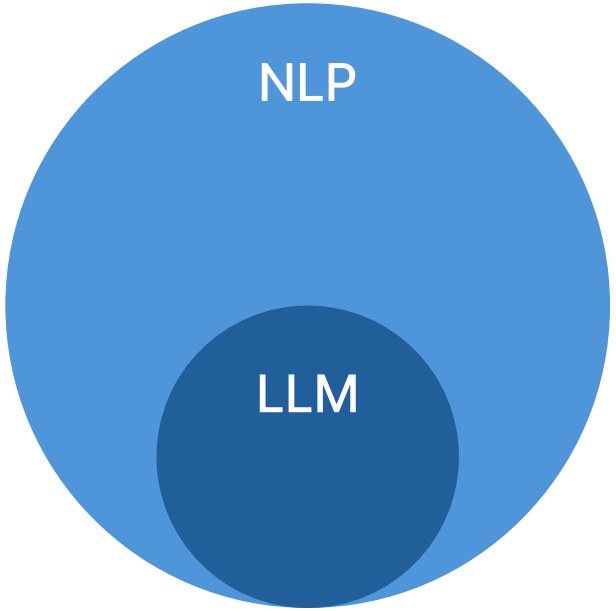
Large Language Model (LLM)

- Language Model의 발전된 형태로, 대규모 데이터 훈련과 복잡한 구조를 통해 다양한 작업을 수행
- LM의 auto-regressive 학습 방식을 그대로 따르지만, 규모의 차이



01 Introduction

NLP vs LLM



Natural Language Processing (NLP)

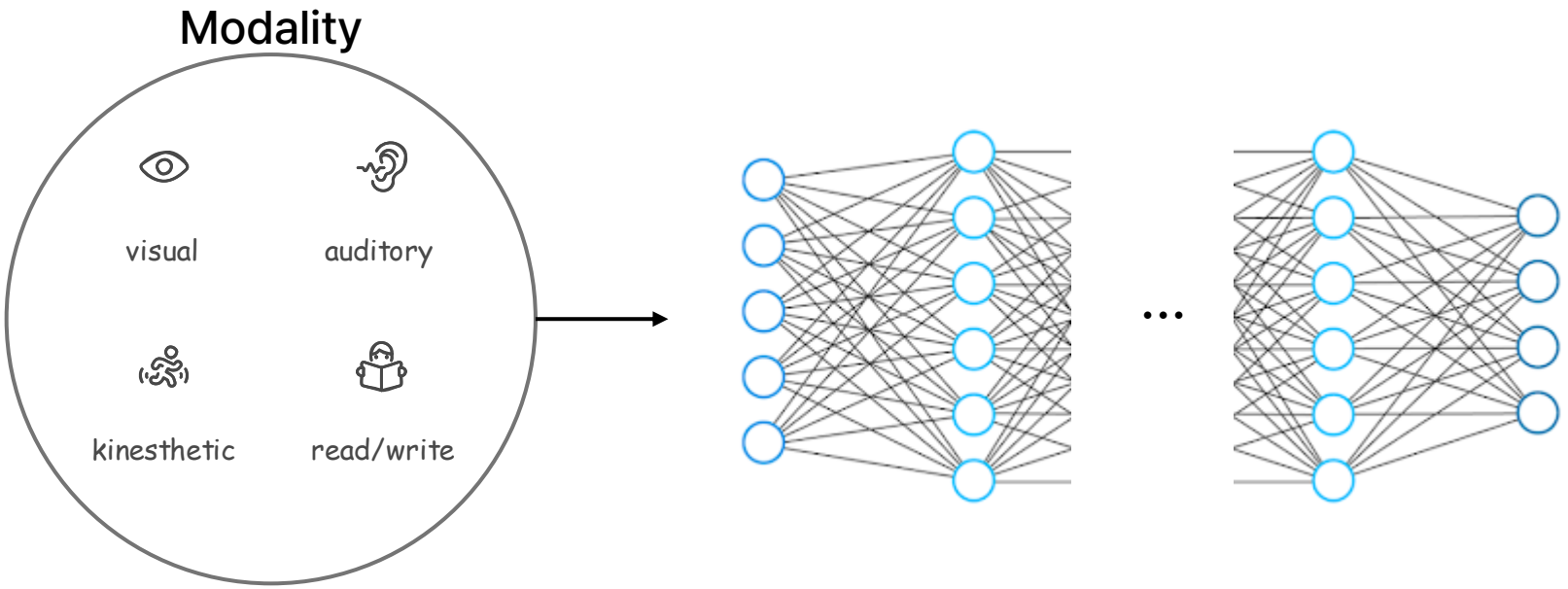
- 컴퓨터가 사람의 언어를 이해하고 처리할 수 있도록 하는 연구 분야

Large Language Model (LLM)

- NLP의 한 부분으로, 대량의 텍스트 데이터를 기반으로 훈련된 언어 모델

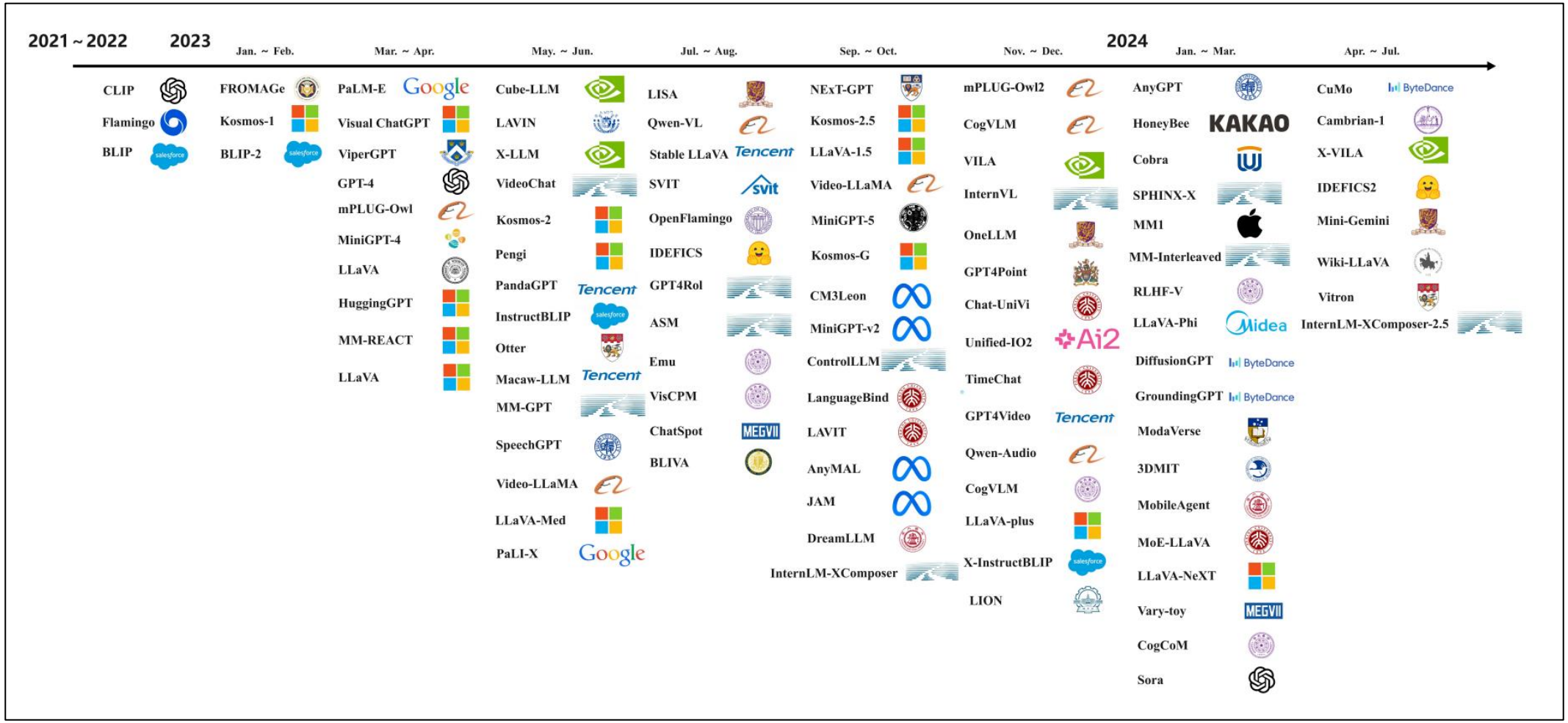
Multimodal LLM

- 텍스트를 넘어 이미지, 오디오 등 다른 modality 정보를 이해하고 상호작용 할 수 있는 LLM
- 다른 modality와 T 시점까지 주어진 단어들을 기반으로 다음 단어를 예측



02 Multimodal LLM

The timeline of MLLMs

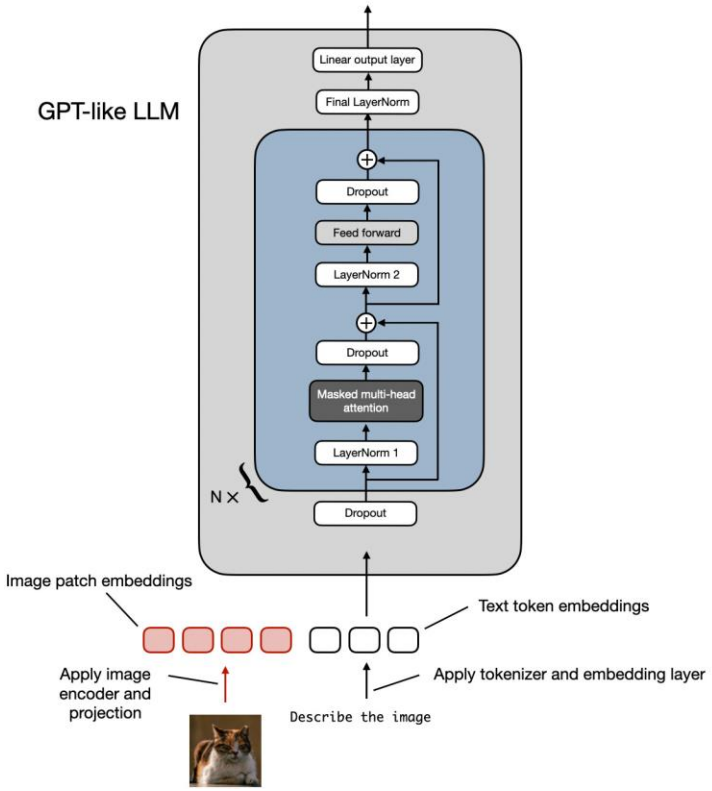


02

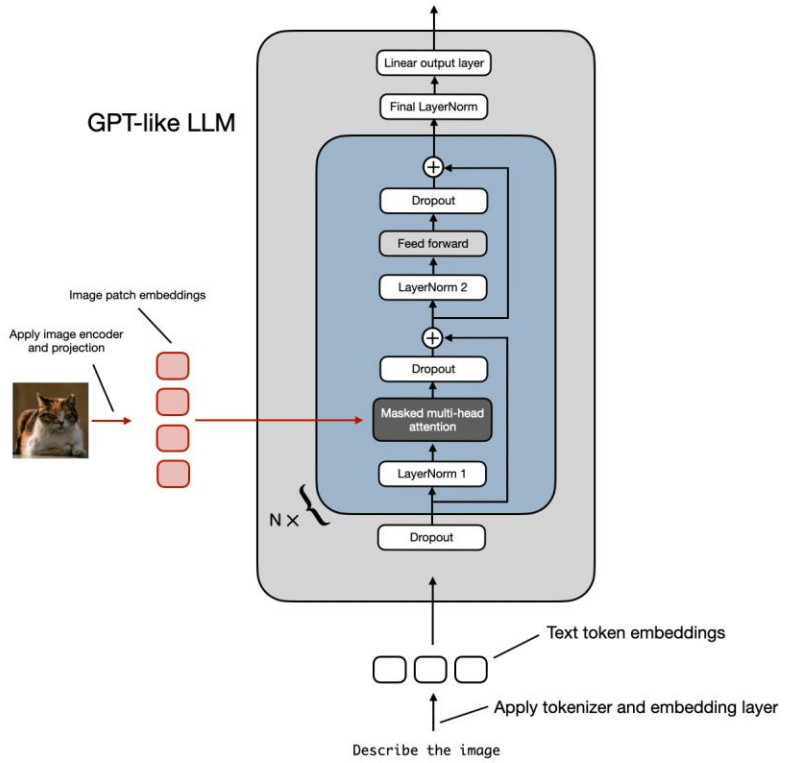
Multimodal LLM

Main approaches to building MLLMs

Method A: Unified Embedding Decoder Architecture



Method B: Cross-Modality Attention Architecture



02

Multimodal LLM

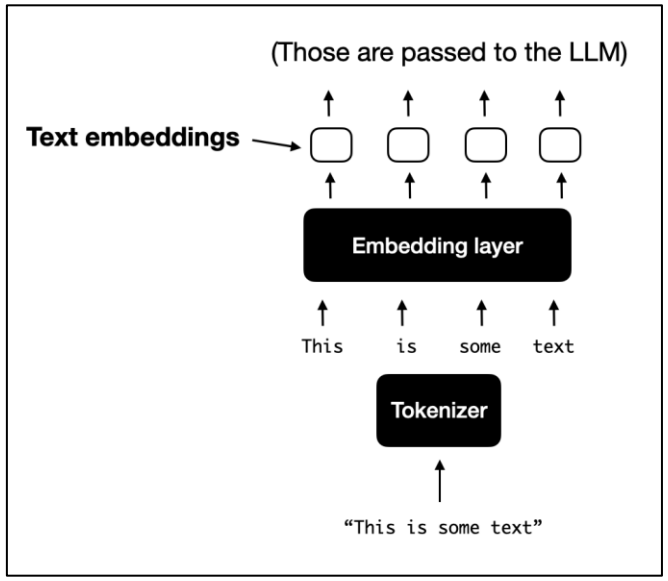
Unified Embedding Decoder Architecture

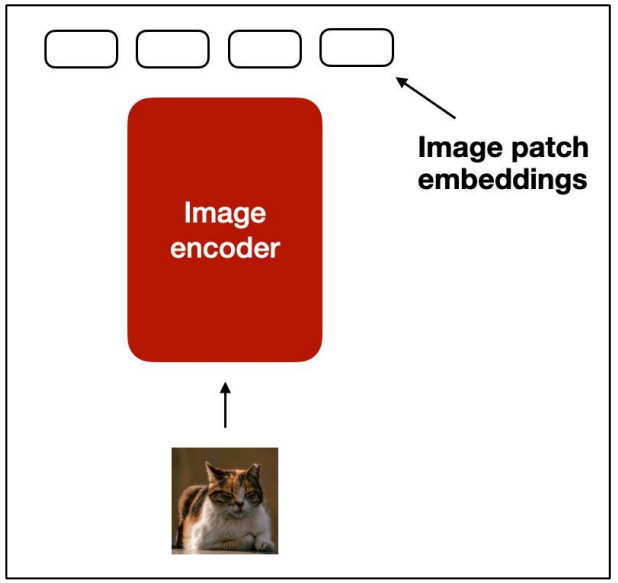
Typical LLM (text-only)

- Tokenizer를 통해 text를 tokenizing
- Embedding layer를 통과함으로써, token embedding vector를 얻음

Image encoder

- Text의 tokenization과 embedding과 유사하게, image encoder module을 통해 image embedding 생성



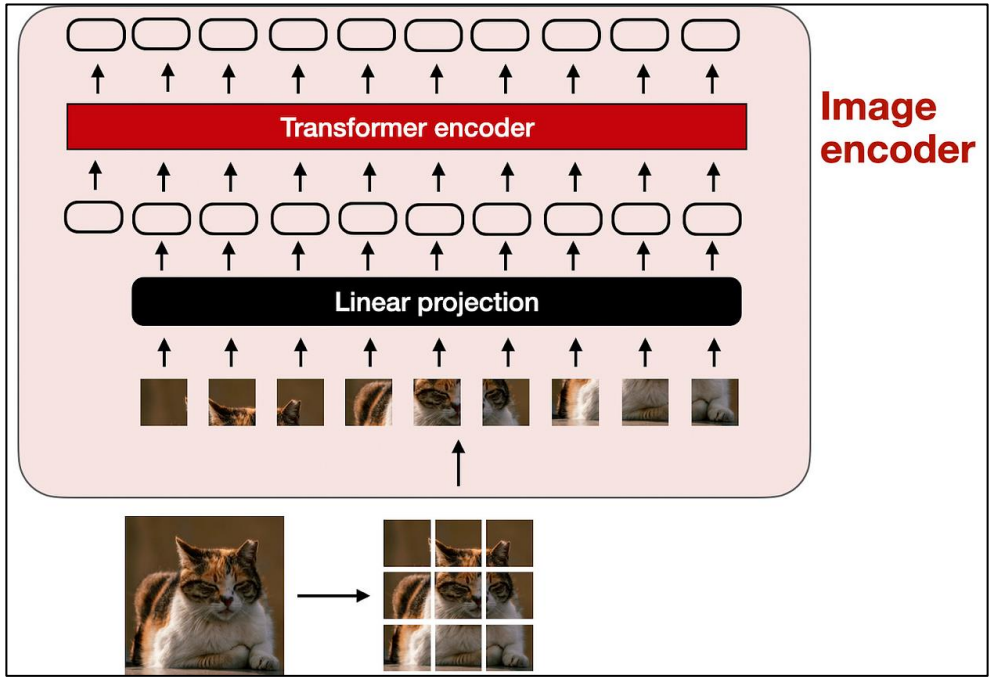


Multimodal LLM

Unified Embedding Decoder Architecture

Image encoder

- 이미지를 처리하기 위해 이미지를 작은 패치들로 분할
- Transformer encoder와 호환되는 embedding size를 만들기 위해 Linear projection 통과
- 예시로 visual encoder를 vision transformer 사용

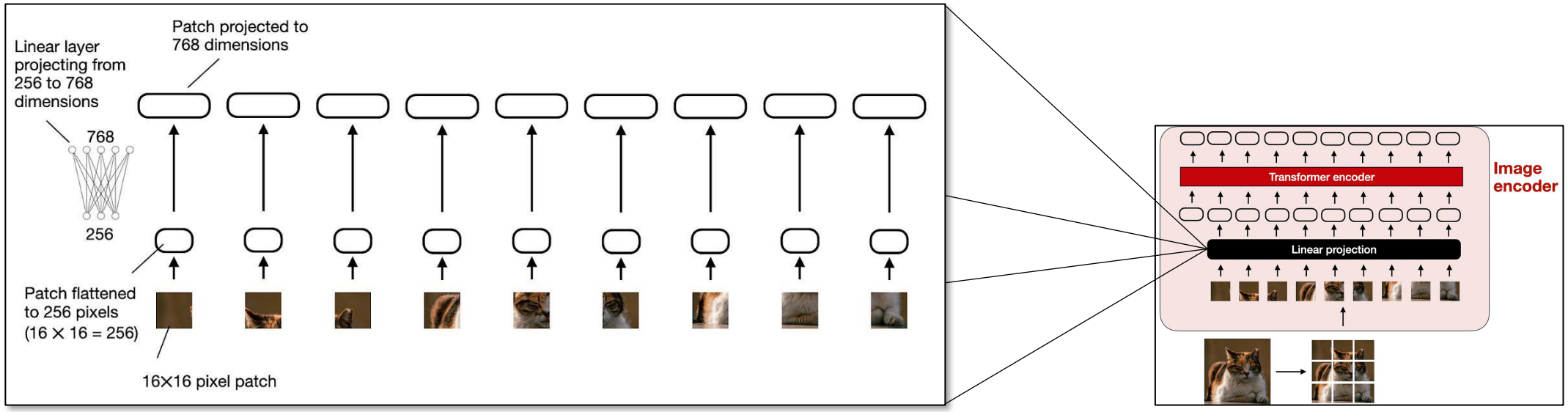


Multimodal LLM

Unified Embedding Decoder Architecture

The role of the linear projection

- Linear projection은 fully connected layer로 구성
- Transformer encoder와 호환되는 embedding size를 만들기 위함
- 16x16 pixel patch라고 가정
- 256-dimensional vector를 768-dimensional vector로 up-projected



02 Multimodal LLM

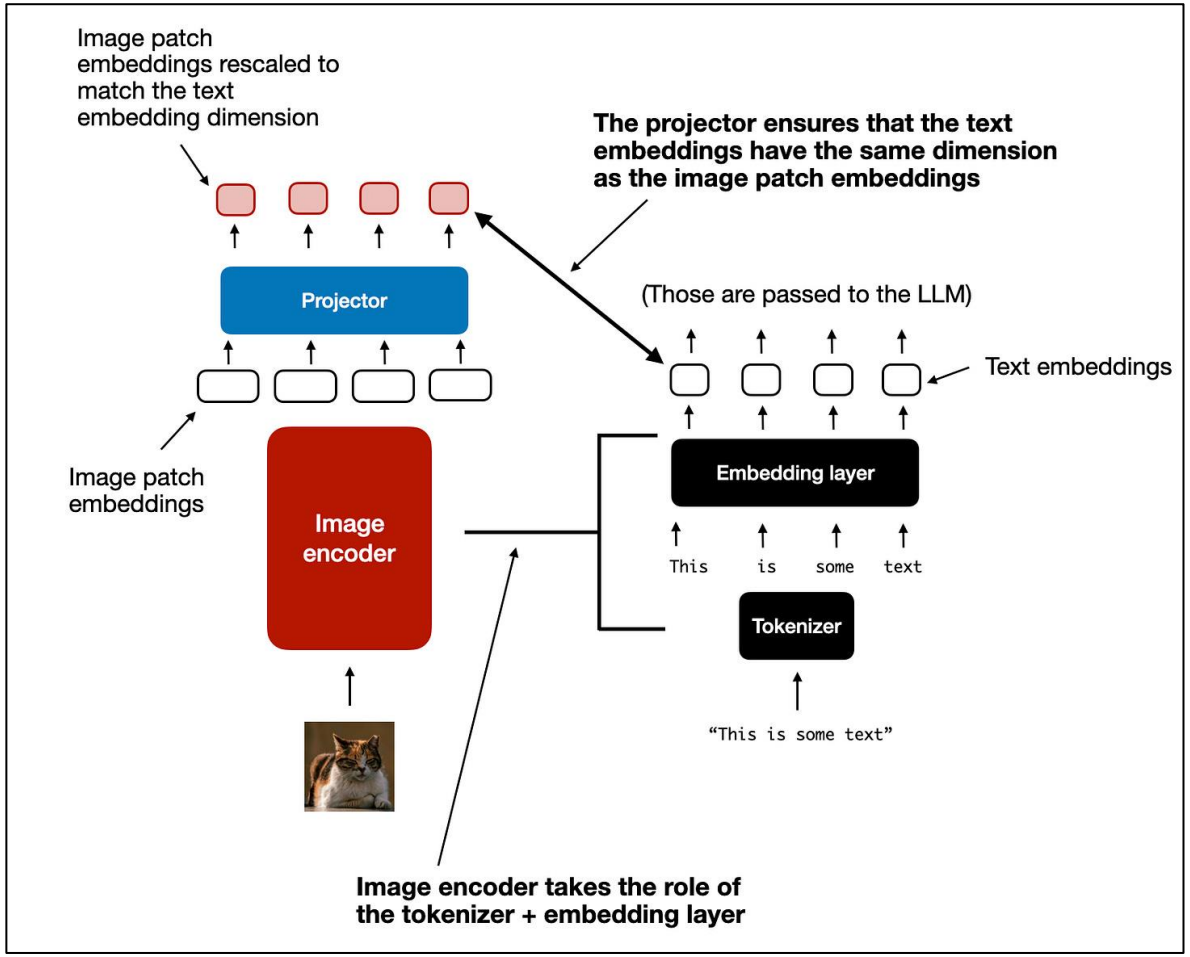
Unified Embedding Decoder Architecture

Projector

- Image encoder의 output 차원과 embedded text token의 차원을 같게 만들기 위함
- Linear projector, Q-Former, Resampler

same embedding dimension

Image patch embeddings == Text token embeddings



02

Multimodal LLM

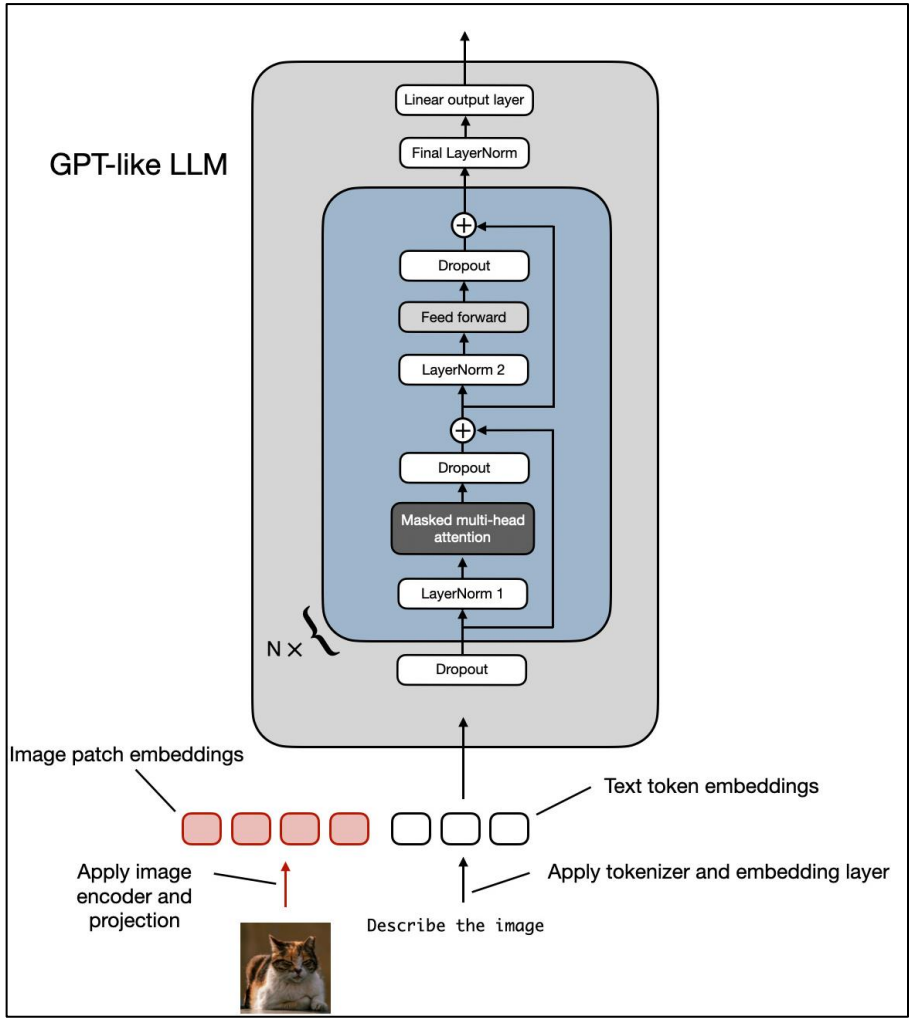
Unified Embedding Decoder Architecture

Projector

- Image encoder의 output 차원과 embedded text token의 차원을 같게 만들기 위함
- Linear projector, Q-Former, Resampler

same embedding dimension

Image patch embeddings == Text token embeddings



02

Multimodal LLM

Unified Embedding Decoder Architecture

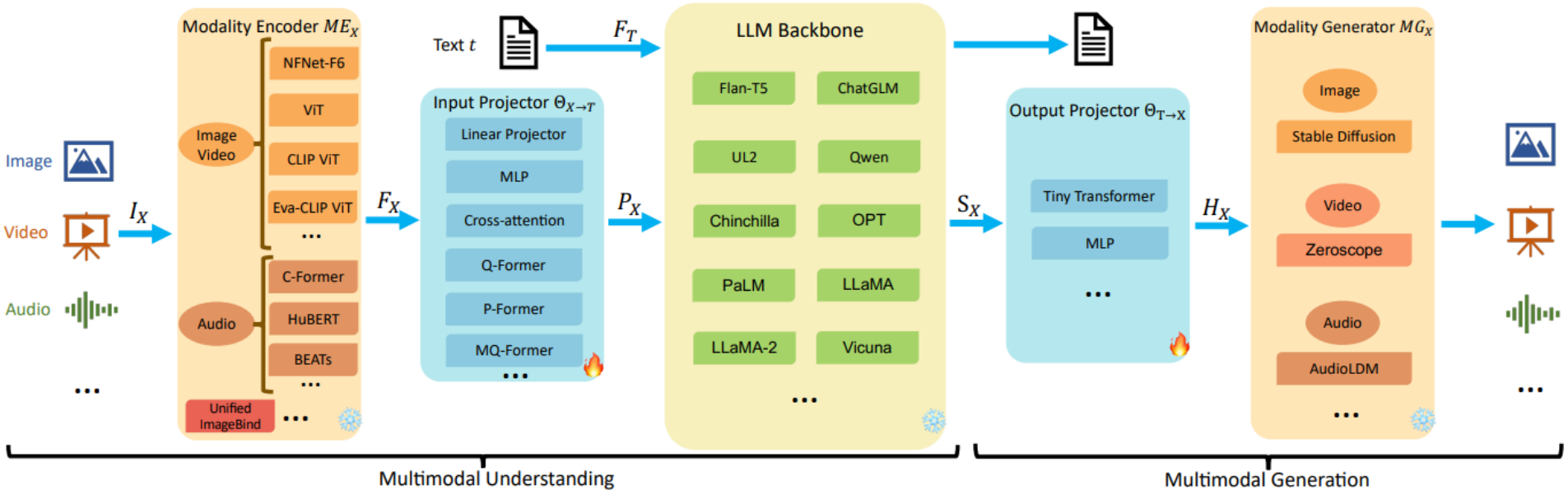


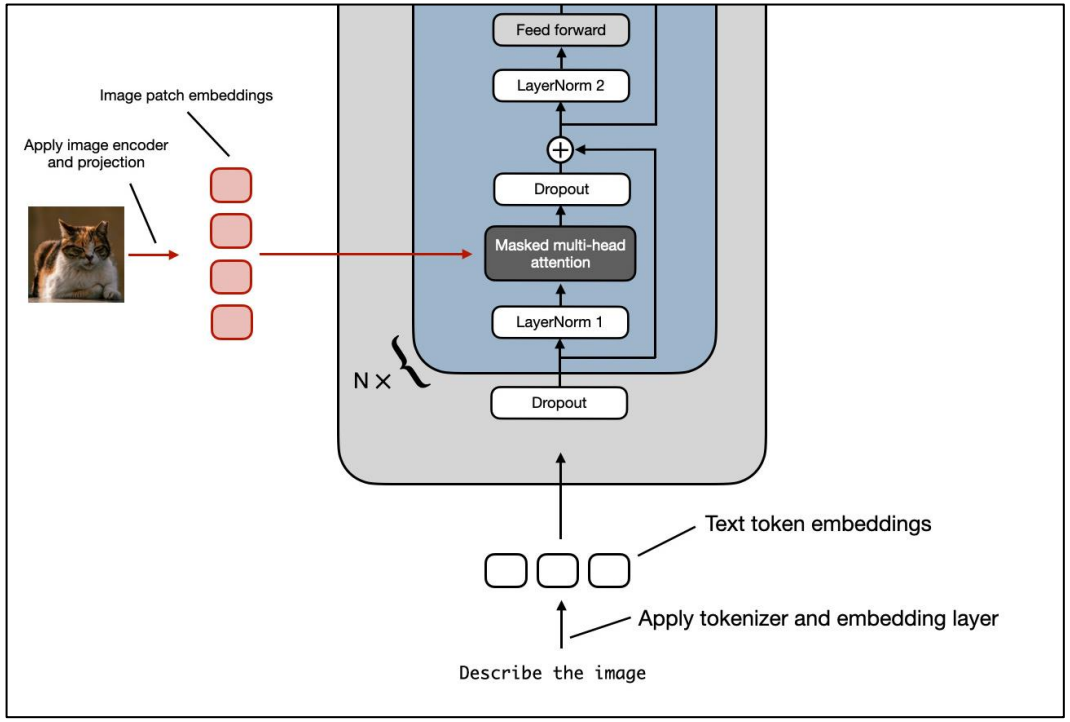
Figure 2: The general model architecture of MM-LLMs and the implementation choices for each component.

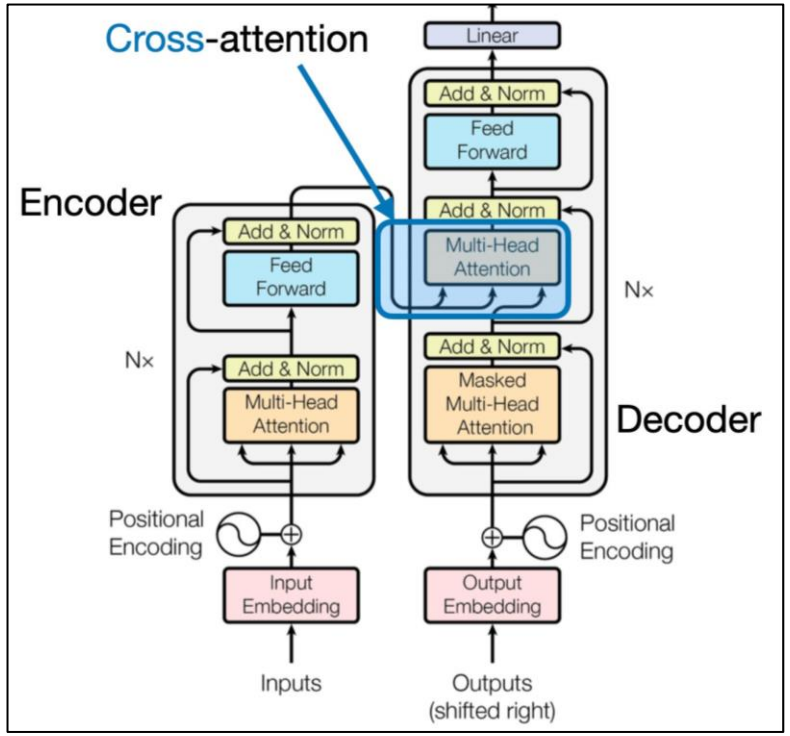
02

Multimodal LLM

Cross-Modality Attention Architecture

- Cross-attention mechanism 활용
- Image patch embedding을 LLM의 Input이 아닌, multi-head attention layer에 연결
- Transformer architecture와 비슷한 구조를 가지고 있고, 동일한 아이디어 적용



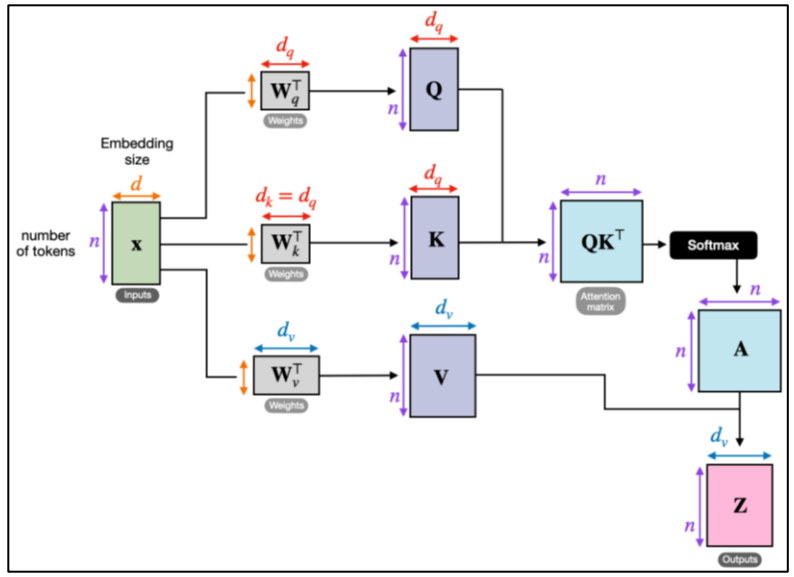


02

Multimodal LLM

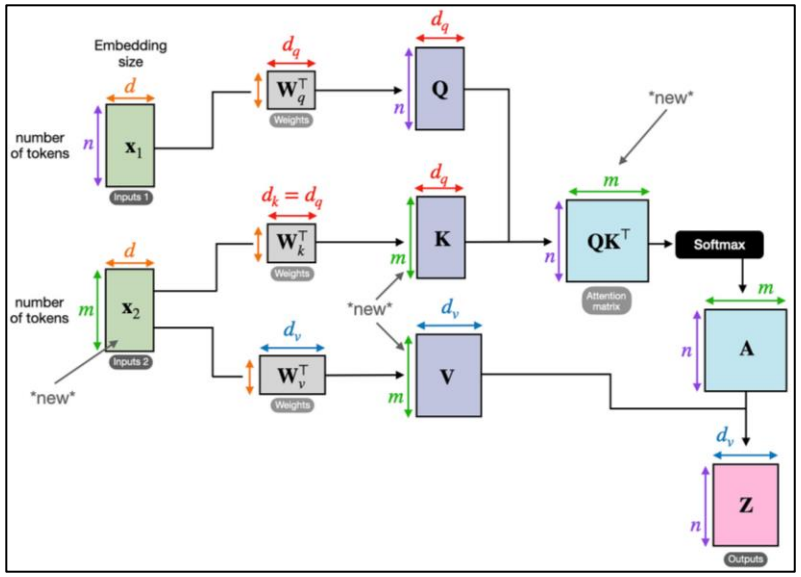
Cross-Modality Attention Architecture

Self-Attention



- 동일한 입력 시퀀스 내에서 관계를 파악할 때 사용
- 각 요소가 다른 모든 요소와의 관계를 고려할 수 있음

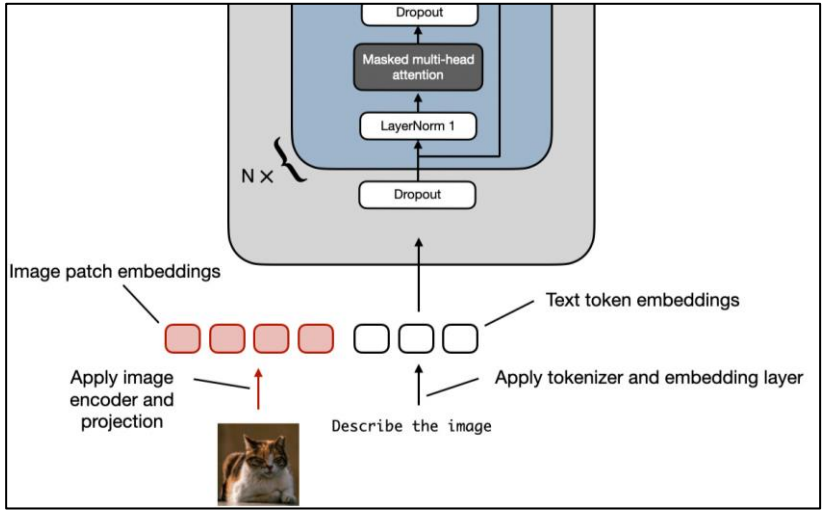
Cross-Attention



- 2개의 서로 다른 입력 시퀀스 간의 관계를 고려
- 일반적으로 Q는 decoder에서 나오고, K와 V는 encoder에서 나옴
- Multimodal LLM 맥락에서, x_2가 Image encoder의 output

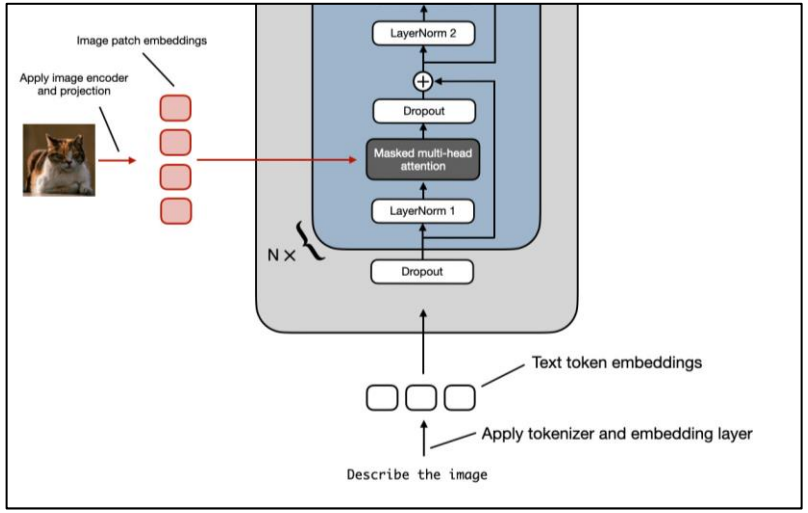
02 Multimodal LLM

A : Unified Embedding Decoder Architecture



- LLM architecture 자체를 수정할 필요가 없음
- 일반적으로 구현하기 쉬움

B : Cross-Modality Attention Architecture



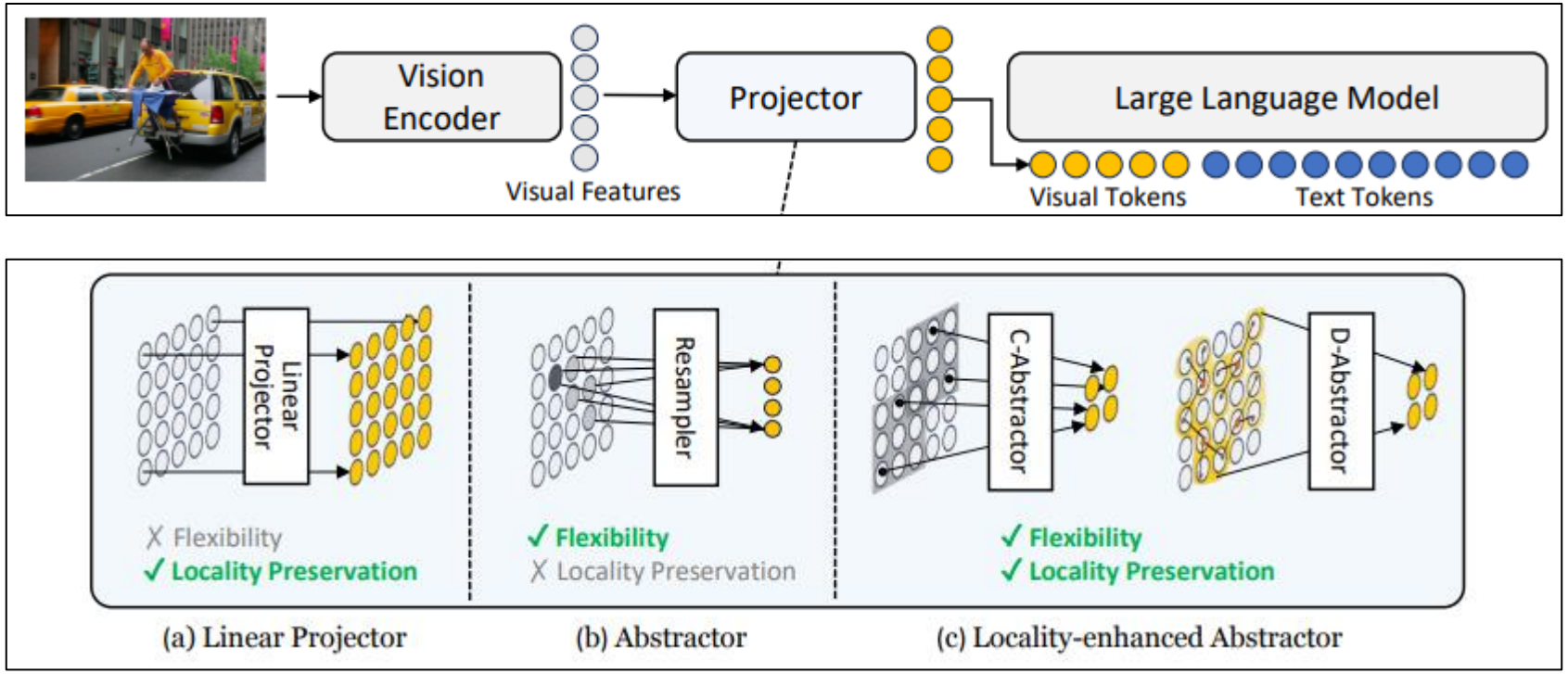
- Image token을 input context에 추가하지 않기에, overload가 되지 않음
- Cross-attention layer에서 추가되기에, 계산적으로 효율적이라 여김
- LLM parameter가 훈련 중에 고정된 경우, 원래 LLM의 text-only 성능을 유지

Multimodal LLM

Multimodal LLMs : Honeybee

Honeybee

- KAKAO에서 2024년 4월에 발표한 Multimodal LLM
- Projector는 encoder와 LLM에 비해 학습 파라미터가 매우 적다는 이유로 중요성이 과소평가 되었다고 판단
- Projector에 집중

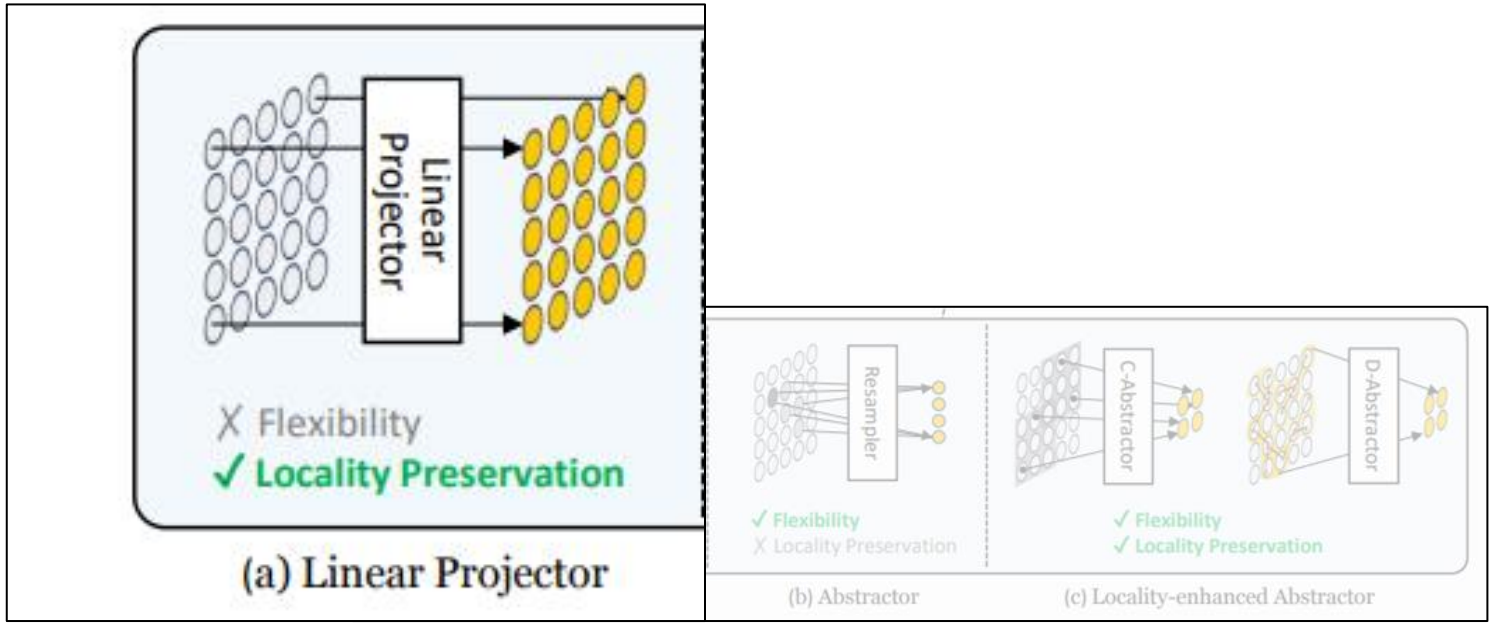


Multimodal LLM

Multimodal LLMs : Honeybee

Linear Projector

- Vision encoder를 통해 나온 visual feature를 하나 혹은 몇 개의 fully connected layer에 통과시킴
- 이를 통해 LLM의 입력 feature dimension과 동일해짐
- 직관적이고 구현이 쉽지만, 모든 visual feature를 LLM의 input으로 넣어야 하는 단점이 있음
- Locality를 보존할 수 있지만, flexibility가 부족



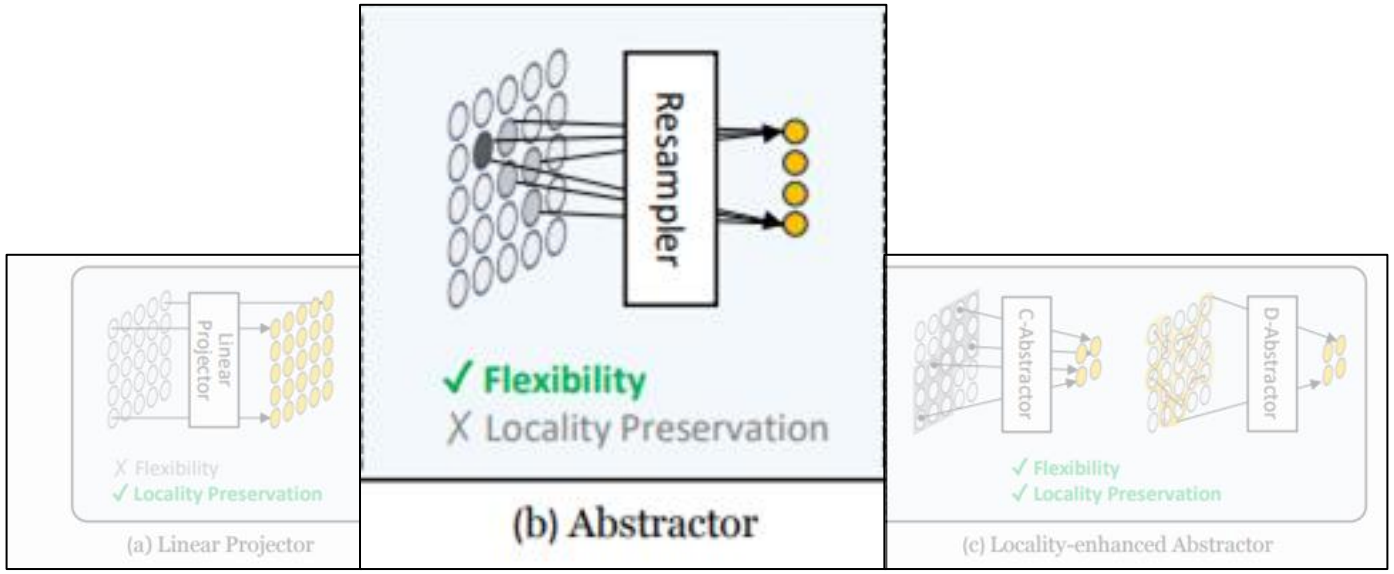
02

Multimodal LLM

Multimodal LLMs : Honeybee

Resampler

- 모델을 학습할 때, 학습자가 지정한 learnable query를 사용
- Visual feature와 learnable query가 cross-attention 형태로 결합
- 아무리 큰 이미지가 오더라도 원하는 길이로 abstraction 된 visual feature를 얻을 수 있음



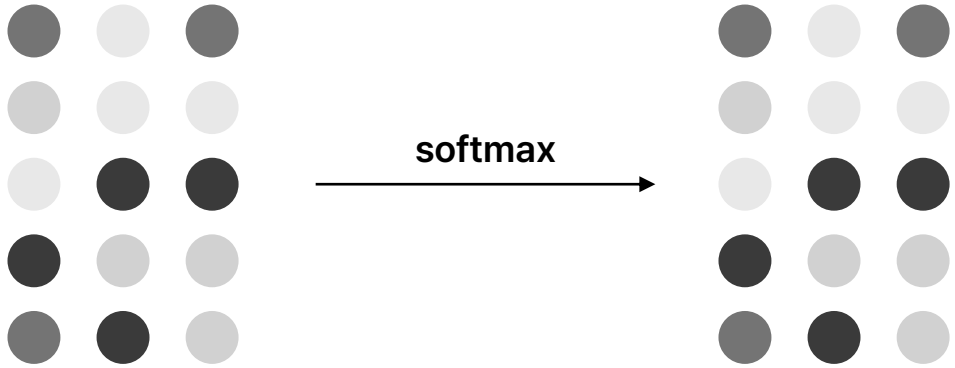
Multimodal LLM

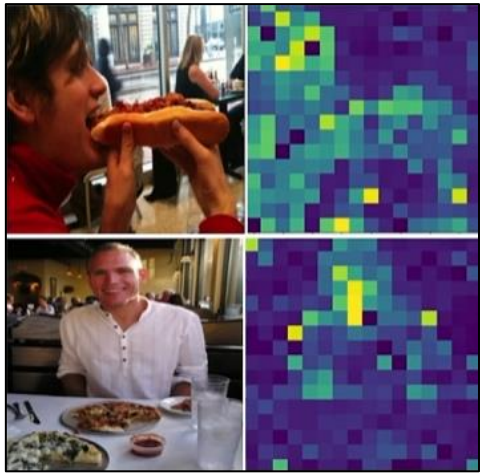
Multimodal LLMs : Honeybee

Resampler

- Cross-attention layer에서 visual feature와 learnable query 사이의 attention map을 계산
- Softmax 함수 특성상 특정 중요 feature의 attention이 강해지기 위해서는, 그 외 다른 feature들의 attention이 낮아져야 함
- Softmax 함수 특성으로 인해, 특정 중요 feature들만 부각된다는 한계가 생김
- 사용자가 attention 값이 낮은 지역의 정보를 알고 싶어 할 때, LLM이 정확한 답변을 주지 못할 가능성이 큼
- Visual projector는 최대한 많은 시각 정보들을 LLM에 넘겨주는 것이 핵심
- Flexibility가 있지만, Locality를 보존할 수 없음

Attention map

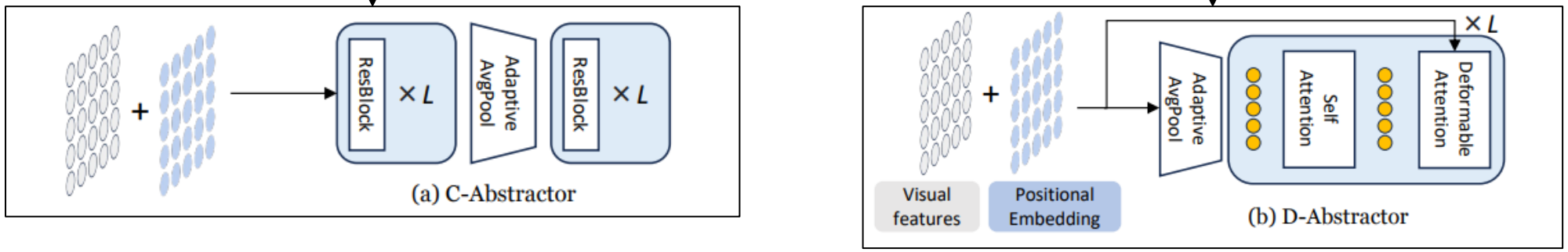




Multimodal LLM

Multimodal LLMs : Honeybee

Visual feature의 수를 효과적으로 제어할 수 있으면서도
손실되는 정보량을 최소화할 수 있는 projector를 고민



C-Abstractor (Convolutional Abstractor)

- Convolution-based
- Local context를 잘 포착하는 convolution 이용
- ResNet Block을 여러 개 사용하여 visual token을 추출

D-Abstractor (Deformable attention-based Abstractor)

- Deformable attention-based
- Locality-awareness 향상
- Abstraction 중에 flexibility 유지

Multimodal LLM

Multimodal LLMs : Honeybee

	Projector	M	s/step	MME POS	MME SR	MMB OL	MMB PR	SEED SR	SEED IL	Avg^N
B1	Linear	144	-	*Unavailable due to inflexiblity*						-
B2	Resampler	144	2.28	75.0	22.2	43.2	62.5	47.5	50.6	43.9
B3	C-Abstractor	144	2.23	135.0	24.4	54.3	66.7	49.0	58.8	53.5
B4	Linear	256	3.04	140.0	24.4	40.7	70.8	48.9	60.9	52.6
B5	Resampler	256	3.12	73.3	24.4	37.0	79.2	44.4	51.8	45.6
B6	C-Abstractor	256	3.07	136.7	26.7	55.6	75.0	52.7	59.3	56.3

- 224x224 입력 이미지
- Spatial relationship task로 성능 비교

	Projector	M	s/step	MME POS	MME SR	MMB OL	MMB PR	SEED SR	SEED IL	Avg^N
B3	C-Abstractor	144	2.23	135.0	24.4	54.3	66.7	49.0	58.8	53.5
B4	Linear	256	3.04	140.0	24.4	40.7	70.8	48.9	60.9	52.6

Linear보다 visual feature(M) 수가 적지만 성능이 좋음

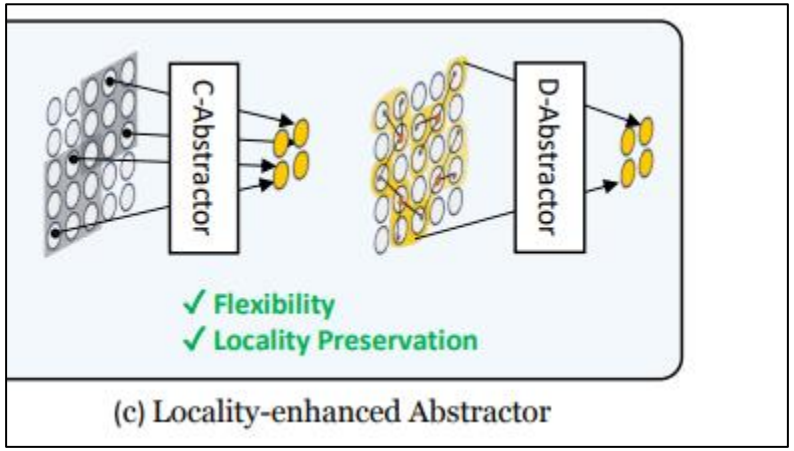
	Projector	M	s/step	MME POS	MME SR	MMB OL	MMB PR	SEED SR	SEED IL	Avg^N
B4	Linear	256	3.04	140.0	24.4	40.7	70.8	48.9	60.9	52.6
B6	C-Abstractor	256	3.07	136.7	26.7	55.6	75.0	52.7	59.3	56.3

Linear와 visual feature(M) 수를 같게 했을 때 성능이 더 좋아짐

Multimodal LLMs : Honeybee

Anomaly detection 적용 가능성

- Honeybee에서 제안한 projector와 기존의 projector의 성능 비교

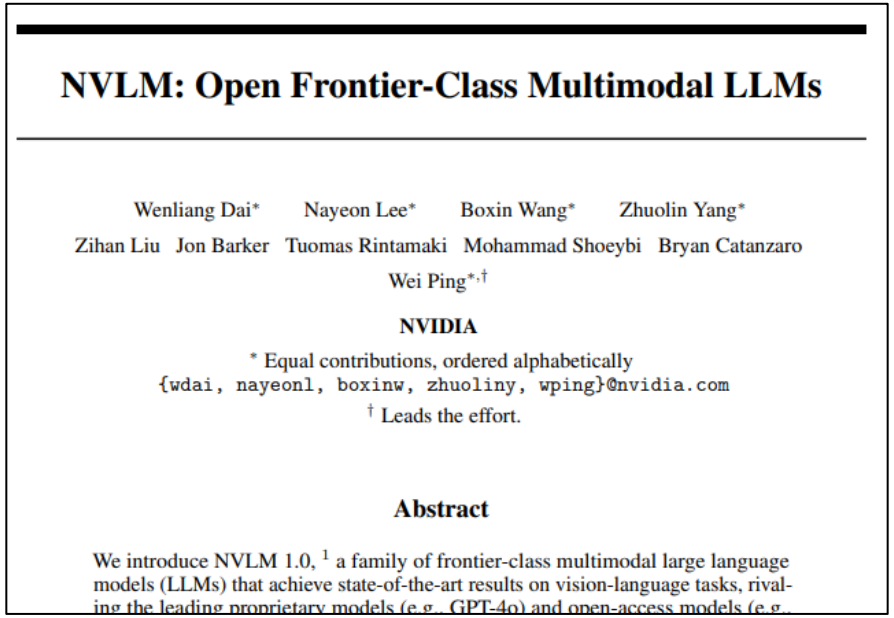


Multimodal LLM

Multimodal LLMs : NVLM

NVLM

- NVIDIA에서 2024년 9월에 발표한 Multimodal LLM
- 앞서 설명한 방법 A(Unified Embedding Decoder Architecture)와 방법 B(Cross-Modality Attention Architecture)를 다룸
- Hybrid approach 방식을 제안하고 3가지 방법을 비교



NVLM: Open Frontier-Class Multimodal LLMs

Wenliang Dai* Nayeon Lee* Boxin Wang* Zhuolin Yang*
Zihan Liu Jon Barker Tuomas Rintamaki Mohammad Shoeybi Bryan Catanzaro
Wei Ping*,†

NVIDIA

* Equal contributions, ordered alphabetically
{wdai, nayeonl, boxinw, zhuoliny, wping}@nvidia.com
† Leads the effort.

Abstract

We introduce NVLM 1.0, [1] a family of frontier-class multimodal large language models (LLMs) that achieve state-of-the-art results on vision-language tasks, rival-
ing the leading proprietary models (e.g., GPT-4o) and open-access models (e.g.,

02

Multimodal LLM

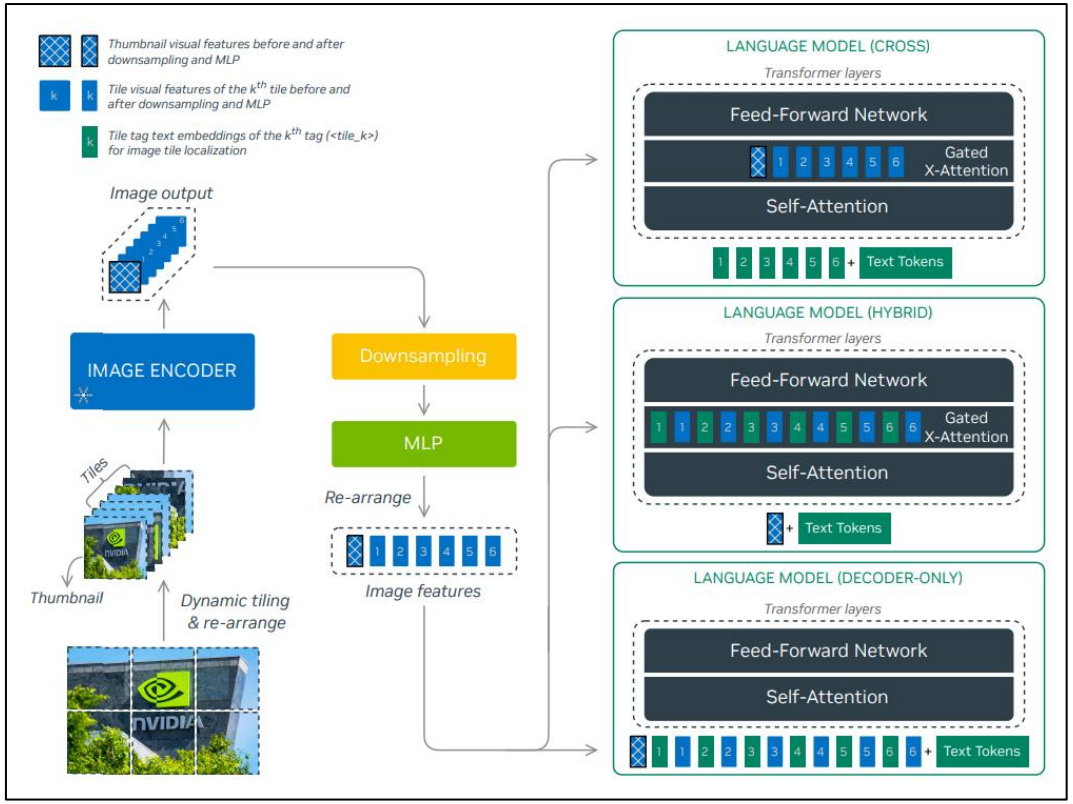
Multimodal LLMs : NVLM

Image encoder

- 일반적인 CLIP 모델 대신 InternViT-6B를 사용
- InternViT-6B는 훈련 중에 frozen 상태를 유지

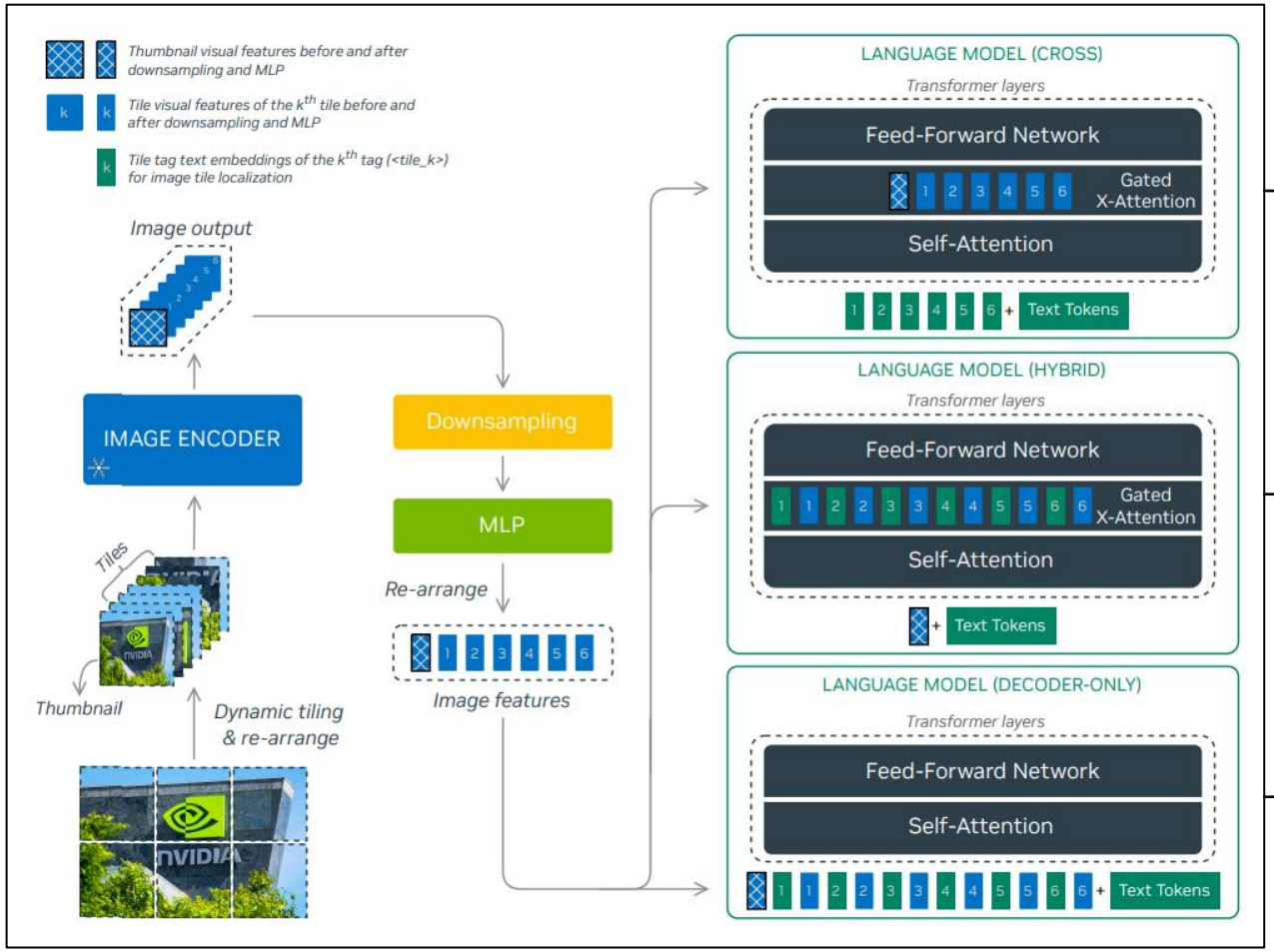
Projector

- Single linear layer가 아닌 multilayer perceptron



02 Multimodal LLM

Multimodal LLMs : NVLM



Method B : Cross-attention based (NVLM-X)

- 고해상도 이미지에서 뛰어난 계산 효율성을 보여줌

Hybrid method (NVLM-H)

- A와 B의 장점을 결합

Method A : Decoder-only (NVLM-D)

- OCR 관련 작업에서 더 높은 정확도를 달성

03

References

[1] Honeybee: Locality-enhanced Projector for Multimodal LLM

[2] NVLM: Open Frontier-Class Multimodal LLMs

[3] A Survey on Multimodal Large Language Models

[4] MM-LLMs: Recent Advances in MultiModal Large Language Models

[5] A Comprehensive Survey of Multimodal Large Language Models: Concept, Application and Safety

[6] Can Multimodal Large Language Models be Guided to Improve Industrial Anomaly Detection?

감사합니다