이제는 AI가 읽고(Language), 보고(Vision), 생성하는 Large-scale Multimodal의 시대입니다

전동현
NAVER Search

NAVER DEVIEW 2023

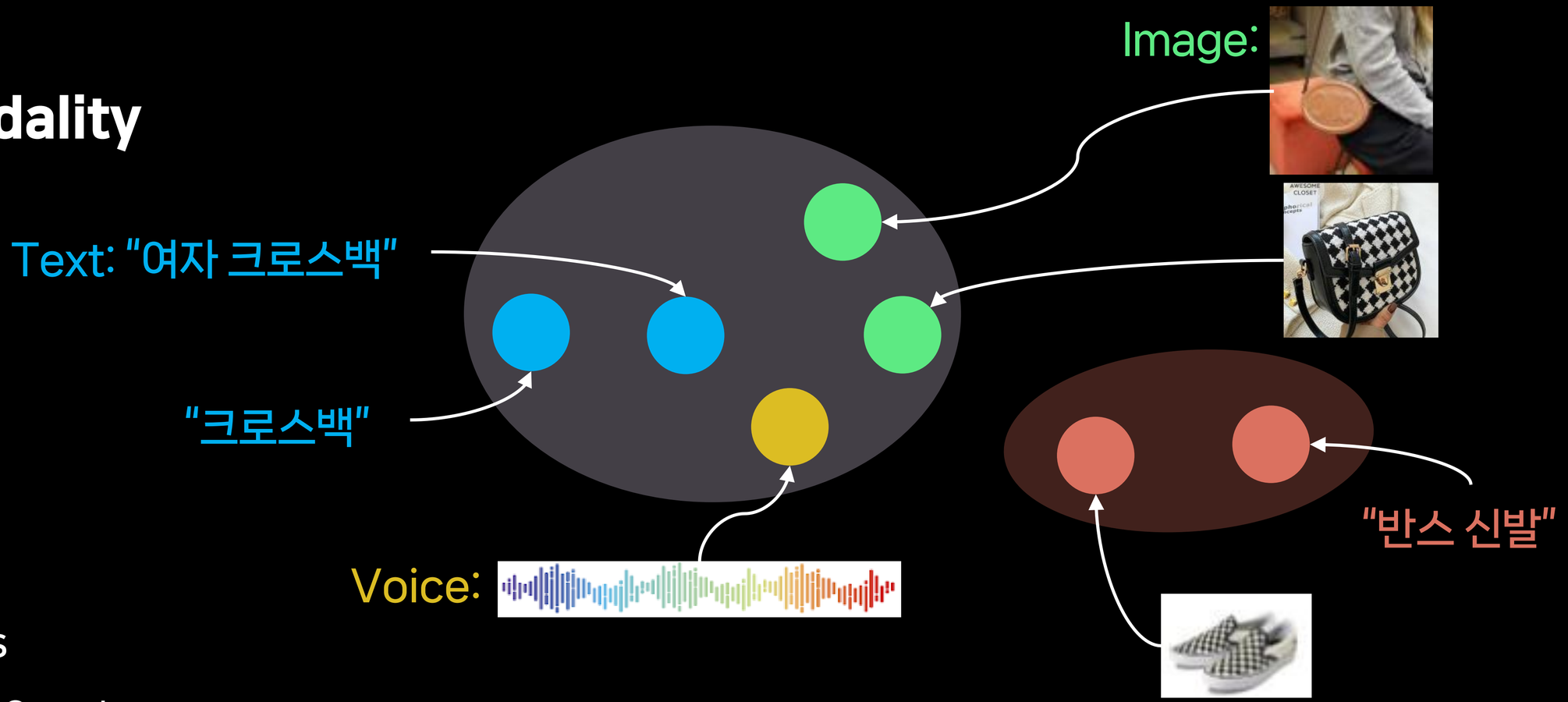
CONTENTS

1. Why Large-scale Multimodal?

1.1 Multimodal

NAVER
DEVIEW
2023

Multimodality



Applications

- Multimodal Search
- Zero-shot Classification
- Multimodal Generation

1.2 Large-scale Foundation Models

NAVER
DEVIEW
2023

Language Models

- Encoder: BERT(110M size, 3.3B words), RoBERTa, ELECTRA, …
- Encoder-Decoder: BART, T5, UL2 (20B), …
- Decoder: GPT3 (175B size, 300B tokens), PaLM (540B size, 780B tokens)

	LaMDA	BlenderBot 3	Sparrow	ChatGPT/ InstructGPT	Assistant
Org	Google	Meta	DeepMind	OpenAI	Anthropic
Access	Closed	Open	Closed	Limited	Closed
Size	137B	175B	70B	175B	52B
Pre-trained Base model	Unknown	OPT	Chinchilla	GPT-3.5	Unknown
Pre-training corpora size (# tokens)	2.81T	180B	1.4T	Unknown	400B

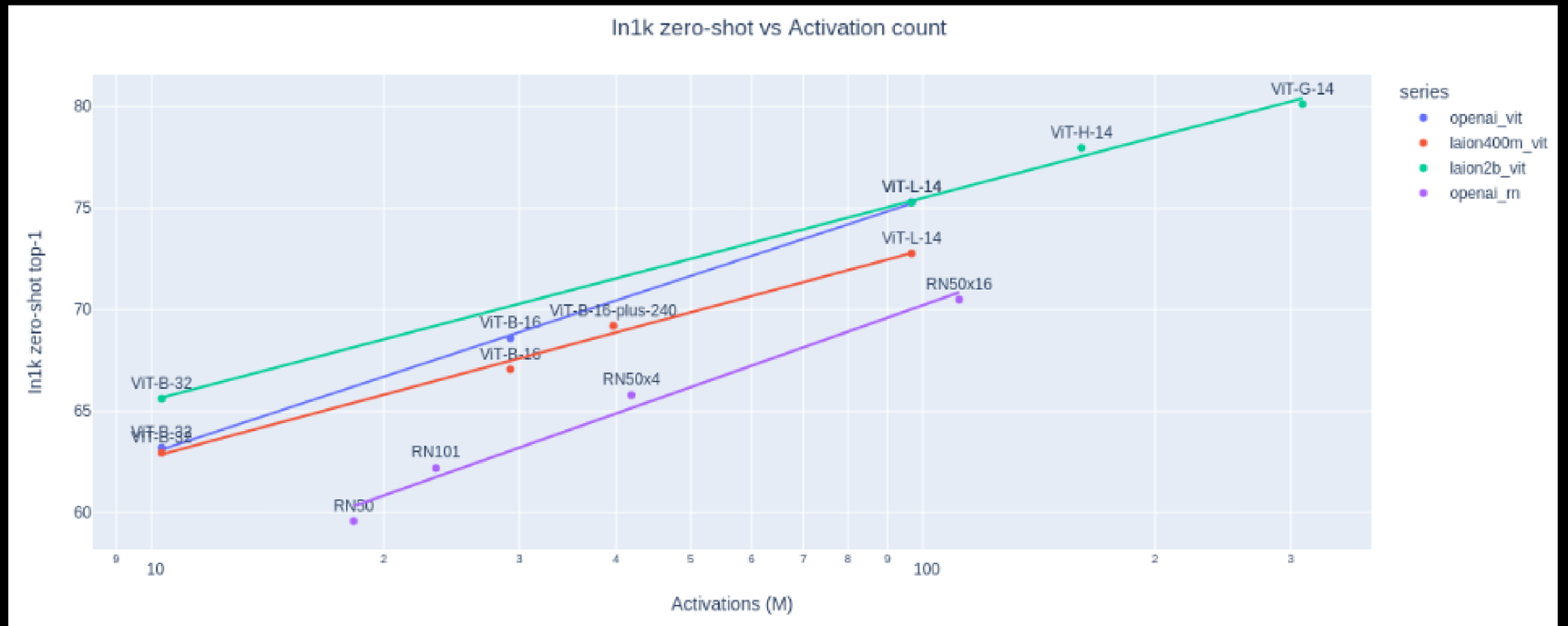
https://huggingface.co/blog/dialog-agents

1.2 Large-scale Foundation Models

NAVER
DEVIEW
2023

Vision-Language Models

- Encoder: CLIP (150M size, 400M image-text pair), ALIGN (1.8B i-t pair), Florence (900M i-t pair), ⋯
- Image-to-text Decoder: SimVLM, Flamingo, ⋯
- Encoder+Decoder: BLIP, BEiT-3, CoCa (2.1B size, 4.8B i-t pair), PaLI (17B size, 10B i-t pair)



https://laion.ai/blog/large-openclip

1.2 Large-scale Foundation Models

NAVER
DEVIEW
2023

Text-to-Image Generation

- DALL-E (1.2B size, 250M image-text pair), DALL-E2(3.5B size), Parti(20B size), …

A green sign that says "Very Deep Learning" and is at the edge of the Grand Canyon. Puffy white clouds are in the sky.

350M

750M

3B

20B

https://parti.research.google

1.3 English Image-Text Dataset

NAVER
DEVIEW
2023

공개되어 있는 Large-scale multimodal dataset

1. LAION−400M/5B[1] : Common Crawl 웹데이터에서 image와 alt-text를 수집

LAION-2B-en: 2.32B

LAION-2B-multi: 2.26B, 다국어

LAION-High-Resolution: 170M, 고화질 이미지 (>=1024)

LAION-Aesthetic: 120M, 미술 이미지

[1] https://laion.ai/blog/laion-5b/

1.3 English Image-Text Dataset

NAVER
DEVIEW
2023

공개되어 있는 Large-scale multimodal dataset

2. Public Multimodal Dataset—70M[2]

- Conceptual Captions (CC3M, CC12M)
- WIT (Wikipedia-based Image Text)
- Localized Narratives
- RedCaps
- COCO
- SBU Captions
- Visual Genome
- subset of YFCC100M

3. COYO—700M[3]

[2] https://huggingface.co/datasets/facebook/pmd
[3] https://github.com/kakaobrain/coyo-dataset

2. Billion-scale
Image-Text Korean 데이터 구축

2.1 Image-Text Dataset

NAVER
DEVIEW
2023

데이터 소스 종류

- 네이버 내부에서 모은 image-text 데이터 (>1.5B)
- 사용이 허용된 외부 데이터셋에서 한글 추출 (>6M)

2.2 Image-Text Dataset Filtering

NAVER
DEVIEW
2023

이미지 필터링

1. 이미지 메타정보 기반

- image size, image size ratio, white pixel ratio

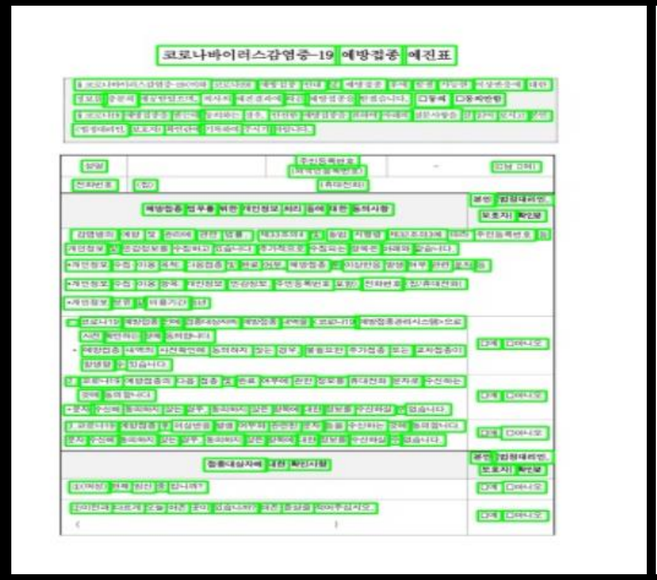
  min(w, h) >= 200

  max(w, h) / min(w, h) <= 3

  num_white_pixels / num_pixels < 0.9

2. OCR

- Detection box 개수, 전체 이미지에서 box 면적의 비율



"ocr_num_box": 42
"ocr_ratio_box": 18%

"watermark_score": 0.9

3. Watermark

- https://github.com/LAION-AI/LAION-5B-WatermarkDetection

2.2 Image-Text Dataset Filtering

NAVER
DEVIEW
2023

쿼리 및 문서 제목 기반 필터링 및 전처리

1. 욕설 / 비속어 / 성인 / 개인정보 포함된 케이스

- 운전면허증 번호, 여권 번호, 주민번호, 전화 번호, 택배 운송장 번호
- 지역 주소, email 주소, URL 주소

2. 무의미한 단순 반복 축약

3. (단독)자모음 / 알파벳 / 특수문자 비율이 50% 이상인 케이스

4. 한글, 영어 외의 언어가 포함된 케이스

5. 로또, 주가, 영수증, 코로나 확진자, 지원금, 악보 등의 count 높지만 학습에 노이즈가 되는 단어들

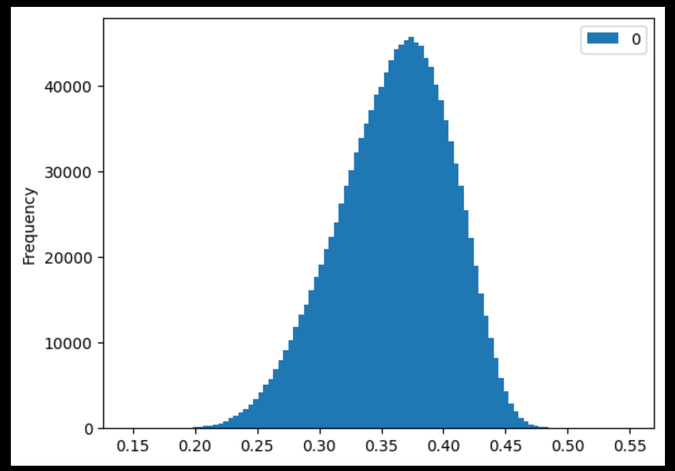
2.2 Image-Text Dataset Filtering

NAVER DEVIEW 2023

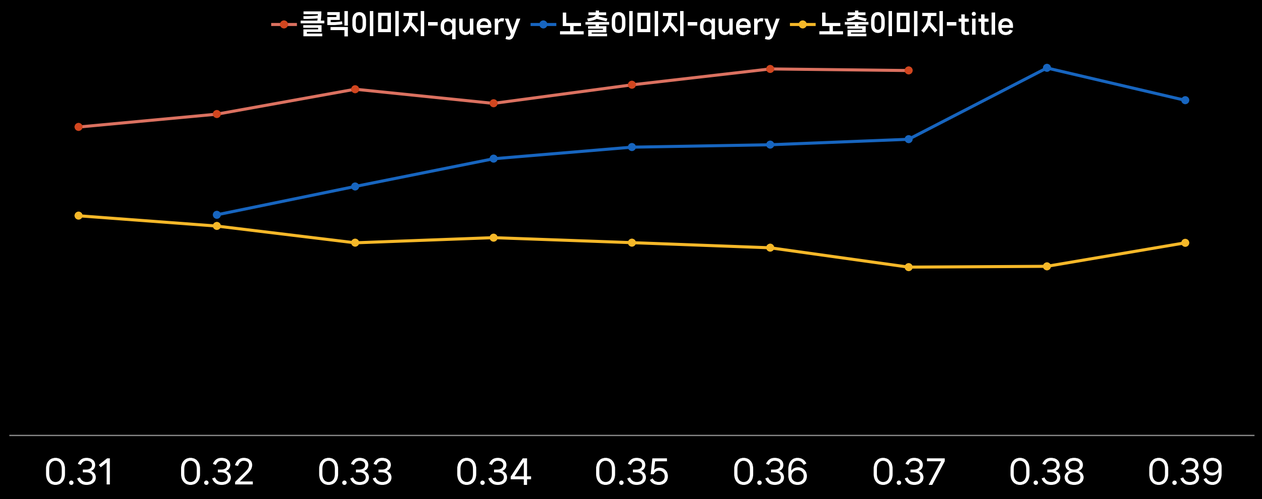
Image-text similarity score (KELIP[1] 활용) 필터링

- 3M 데이터 사이즈에서 zeroshot 성능 비교로 최적의 필터링용 kelip score 찾음
- Text로 문서 제목보다는 쿼리를 많이 사용하는게 더 적절

KELIP score에 따른 데이터 분포



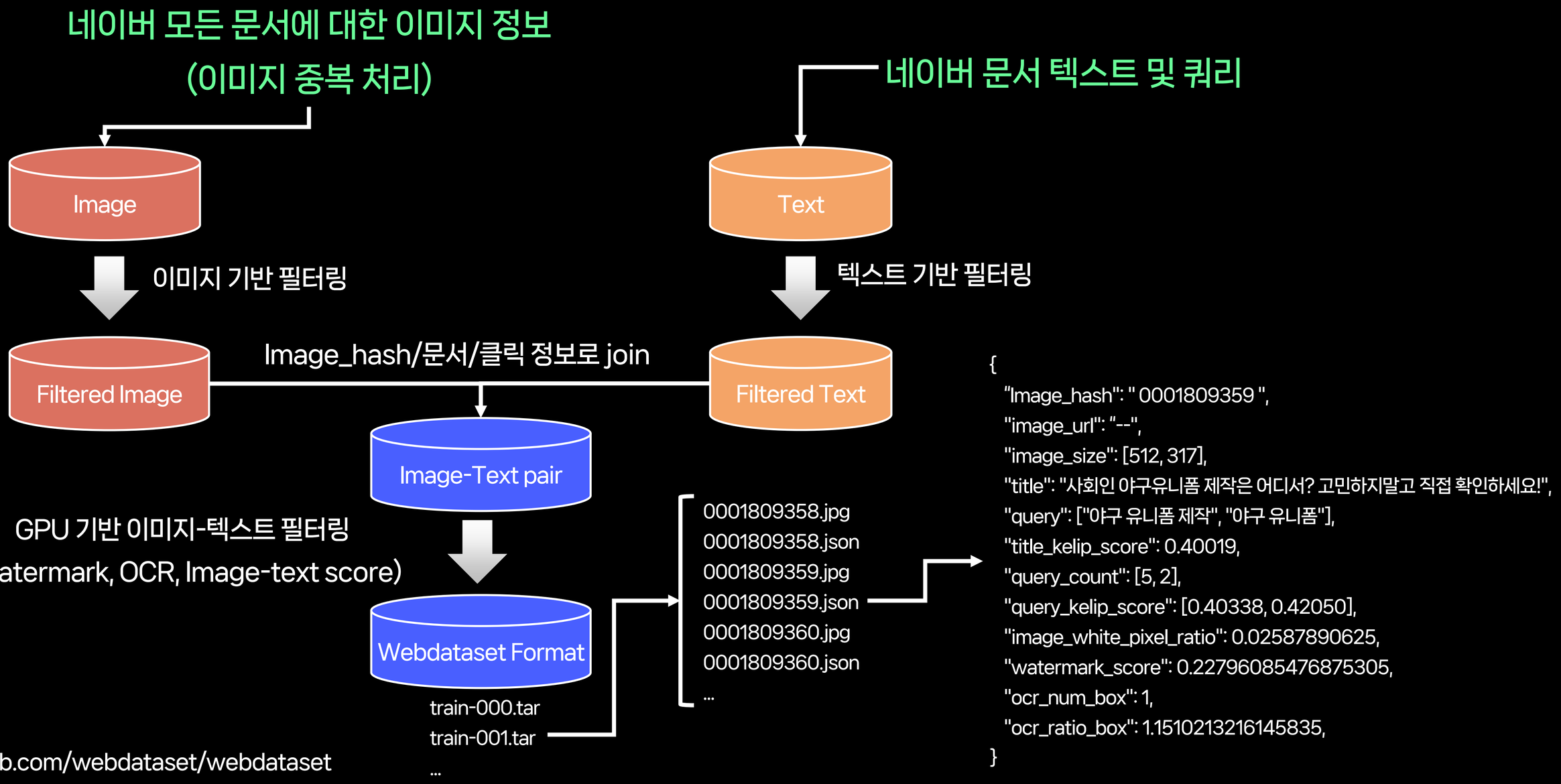
KELIP score에 따른 Imagenet-kr zeroshot 성능



[1] https://github.com/navervision/KELIP

2.3 Image-Query 수집 파이프라인

NAVER DEVIEW 2023



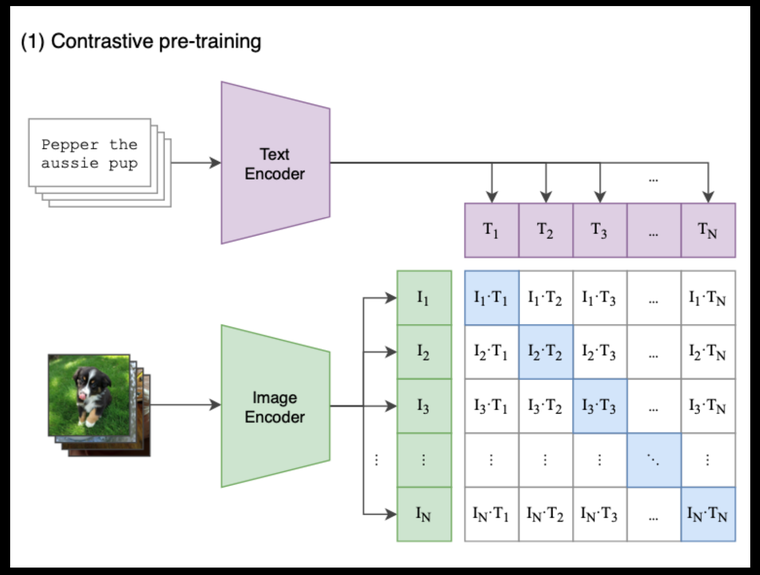
https://github.com/webdataset/webdataset

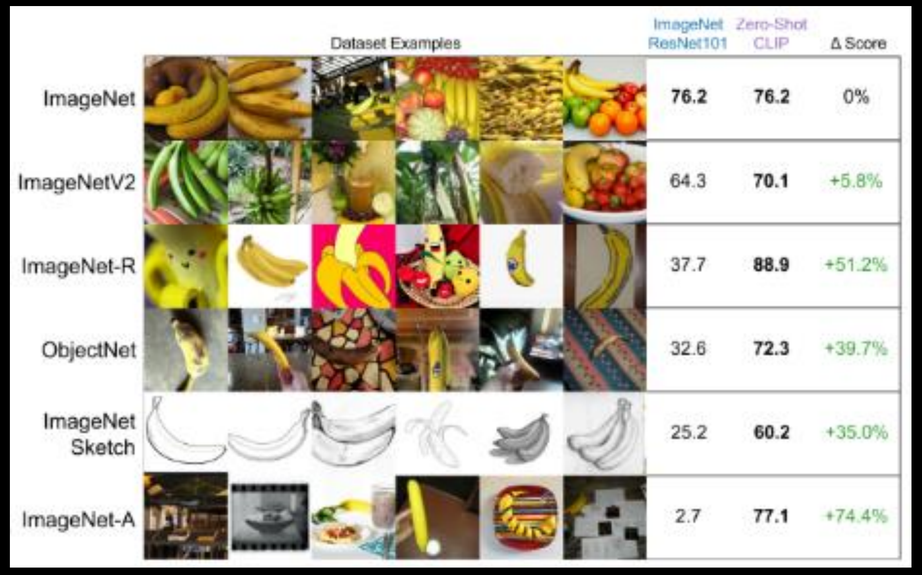
3. Multimodal (Vision-Language) Foundation Modeling

3.1 패션 상품 속성 검색 Model

NAVER
DEVIEW
2023

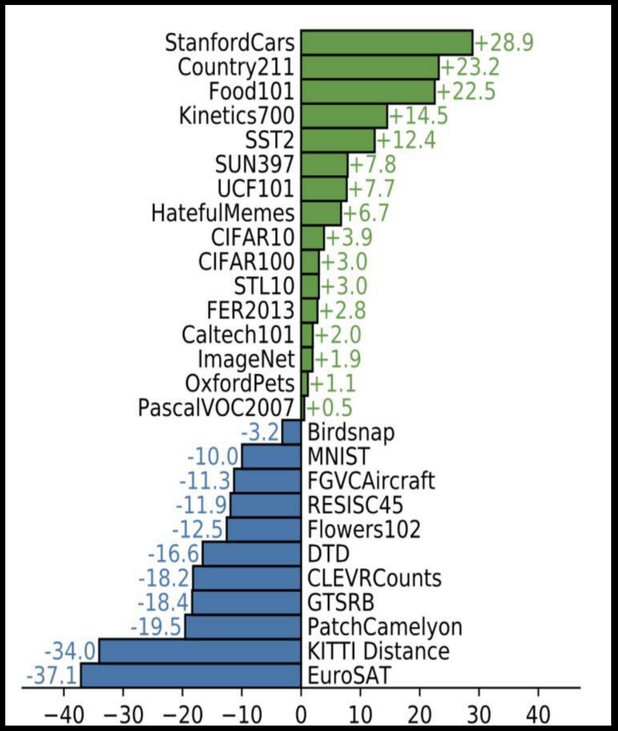
CLIP





CLIP의 한계

- 도메인별 성능 차이가 심해 범용적이지 못함

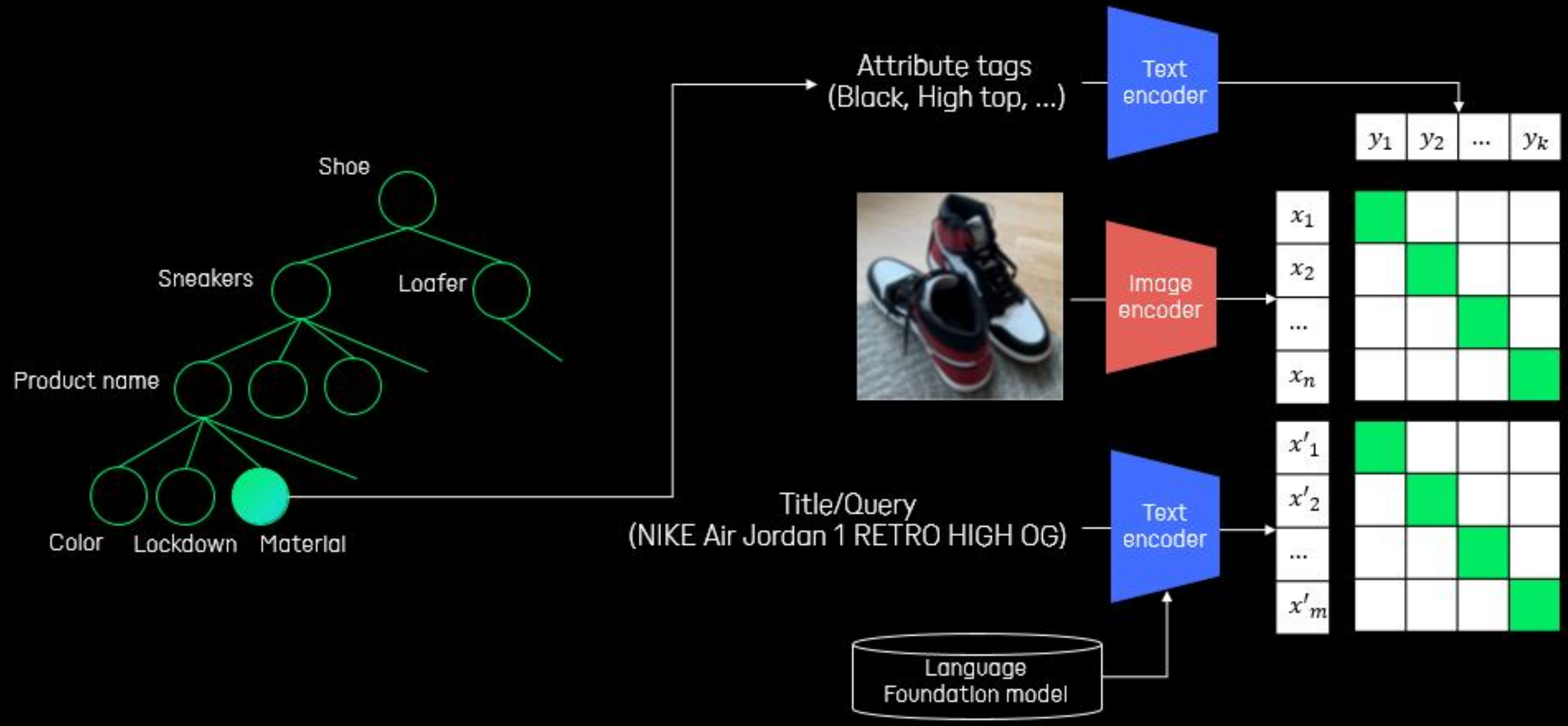


Learning transferable visual models from natural language supervision. PMLR, 2021

3.1 패션 상품 속성 검색 Model

NAVER
DEVIEW
2023

Domain-specific Image-Text Retrieval Model



3.1 패션 상품 속성 검색 Model

NAVER
DEVIEW
2023

Dataset: 쇼핑 패션 데이터 중 Image 와 패션 속성 Text pair, 총 66M

Image+Text(카테고리) <-> Text (속성) 학습

"신발"

속성 키	속성 값
주요특징	리본
주요소재(신발)	스웨이드
토스타일	라운드 토
...	...

속성 키와 속성 값을 template에 넣어 문장화 해서 학습

Ex) "이 상품의 [토스타일]은 [라운드 토] 입니다"

3.1 패션 상품 속성 검색 Model

NAVER
DEVIEW
2023

상품 속성 추출 예시

이미지 Only

query (신발, 원피스, 상의, 치마, 바지...)

	속성 키	속성 값
0	총기장	미니
1	패턴	도트
2	스커트스타일	티어드
3	패턴	레오파드
4	스커트스타일	플리츠/주름

이미지 + '신발'

query (신발, 원피스, 상의, 치마, 바지...)

신발

query: 신발

	속성 키	속성 값
0	주요소재(신발)	메시
1	발목높이	미드탑
2	발목높이	로우탑
3	부가기능	키높이

이미지 + '치마'

query (신발, 원피스, 상의, 치마, 바지...)

치마

query: 치마

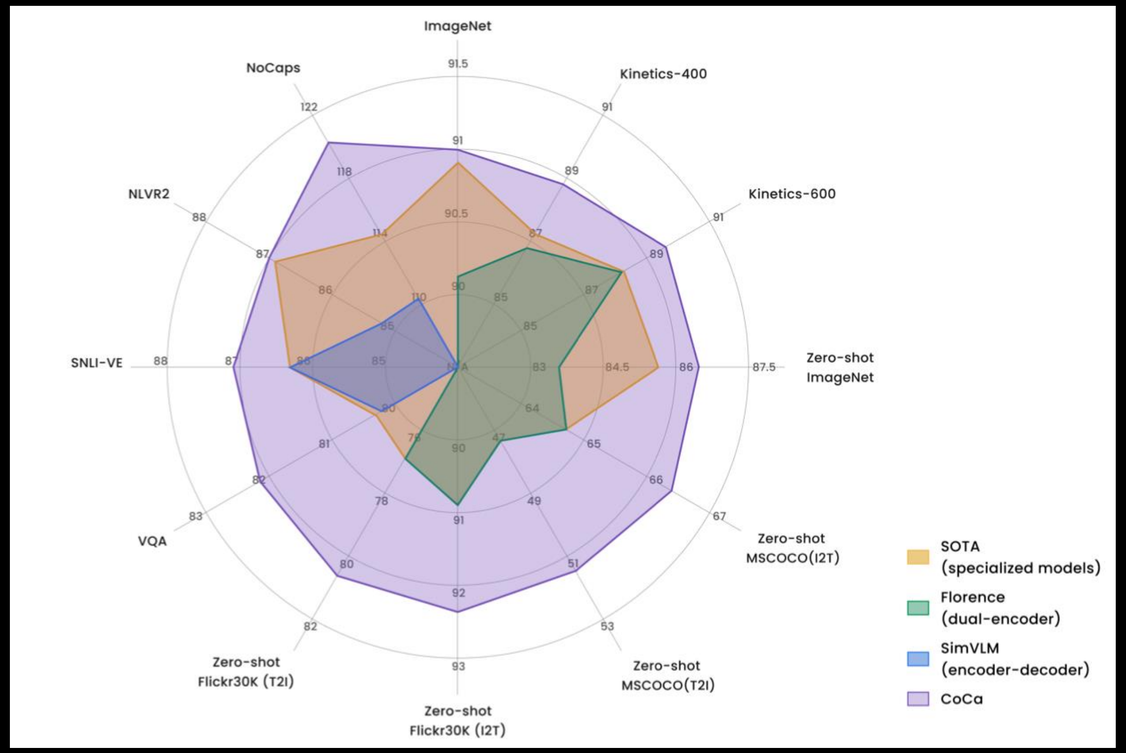
	속성 키	속성 값
0	총기장	미니
1	패턴	도트
2	스커트스타일	플리츠/주름
3	디테일	프릴/러플

3.2 Prompt-CoCa Model

NAVER
DEVIEW
2023

그럼 이제 더 범용적인 모델을 만들어보자

- Contrastive Captioners are Image-Text Foundation Models (CoCa)
- Image/Text Encoder를 활용하는 태스크 뿐만 아니라 Image-to-text 생성도 가능

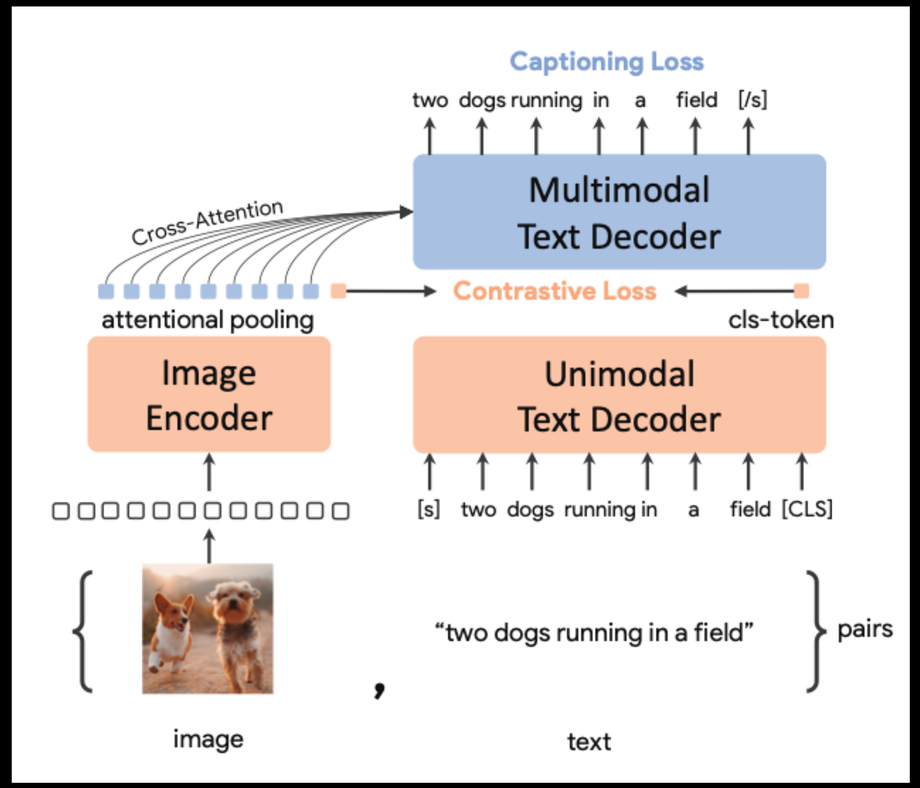


CoCa: Contrastive Captioners are Image-Text Foundation Models, 2022

3.2 Prompt-CoCa Model

NAVER DEVIEW 2023

기존의 CoCa model



Dual-Encoder Contrastive Loss

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Big(\underbrace{\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}}_{\text{image-to-text}} + \underbrace{\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}}_{\text{text-to-image}}\Big),$$

Encoder-Decoder Captioning Loss

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T}\log P_\theta(y_t|y_{<t}, x).$$

최종 Loss

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}}\cdot\mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}}\cdot\mathcal{L}_{\text{Cap}}$$

모델 사이즈

	CoCa-Base	CoCa-Large	CoCa
# Params Image Encoder	86M	303M	1B
# Params Text Decoder	297M	484M	1.1B
# Params Total	**383M**	**787M**	**2.1B**

3.2 Prompt-CoCa Model

NAVER
DEVIEW
2023

문제는 Text 스타일이 너무 다르다!!

문서 제목 스타일

"title": "사회인 야구유니폼 제작은 어디서? 고민하지말고 직접 확인하세요!"

쿼리 스타일

"query": “야구 유니폼 제작“

“query”: "야구 유니폼"

캡션 스타일

"caption": "2021 서울 사회적경제 온라인박람회 포스터"

"caption": "헨리 마티스 작. 브론즈.",

상품명 및 속성 스타일

“product_name": "2021 이브닝 피로연 연주 롱드레스 촬영용드레스",

"category": "패션의류|여성의류|파티복",

"fashion_attribute": {

"주요소재": ["시폰", "폴리에스테르"],

"총기장": "미디",

"소매기장": "민소매",

"종류": "웨딩드레스",

"핏": "슬림핏",

"스커트스타일": "A라인”}

3.2 Prompt-CoCa Model

NAVER
DEVIEW
2023

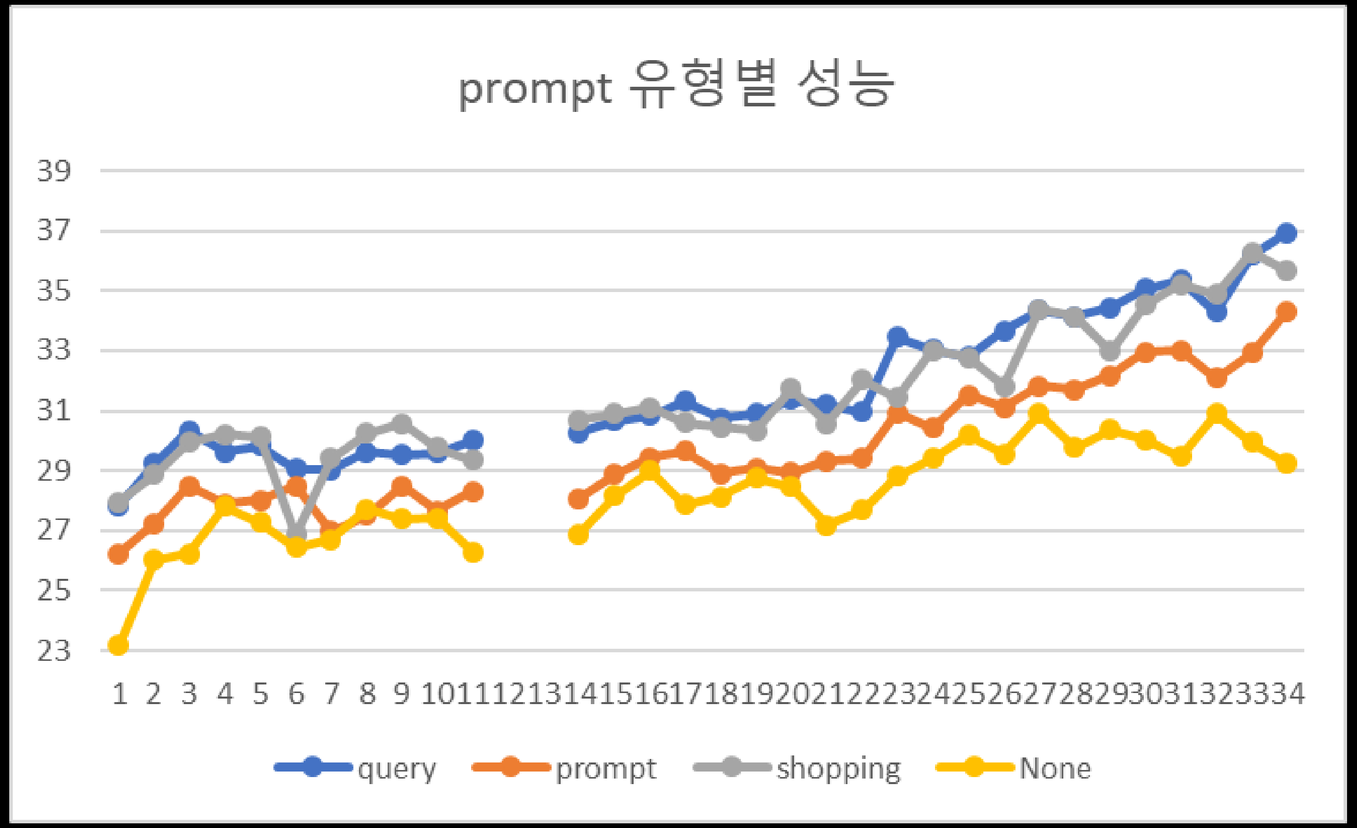
Prompt-CoCa : 데이터 소스별로 Prompt 토큰을 구분해서 학습

- 문서 제목 및 캡션 스타일: “[PROMPT]” + caption/title
- 쿼리 스타일: “[QUERY]” + query
- 상품명 스타일: “[SHOPPING]” + product_name

3.2 Prompt-CoCa Model

NAVER
DEVIEW
2023

Prompt-CoCa 평가: Imagenet-kr zeroshot



3.2 Prompt-CoCa Model

NAVER
DEVIEW
2023

Image-to-text 생성 예시

MSCOCO en/kr Finetuning

En-CoCa: a person riding a skate board at a skate park
Ko-CoCa: 스케이트 보드를 타는 사람이 묘기를 부리고 있다.

En-CoCa: a red and yellow train on the tracks at a station .
Ko-CoCa: 빨간 색과 노란 색의 기차가 기차 역에 들어온다.

En-CoCa: there is a cruise ship in the background .
Ko-CoCa: 한 무리의 사람들이 우산을 들고 해변에 앉아 있다.

En-CoCa: a kitchen with a stove , refrigerator , sink and cabinets .
Ko-CoCa: 나무 캐비닛과 난로가 있는 부엌

이미지-to-쿼리 Finetuning

Ko-CoCa: 샤또 라뚜르

Ko-CoCa: 경복궁 야간개장

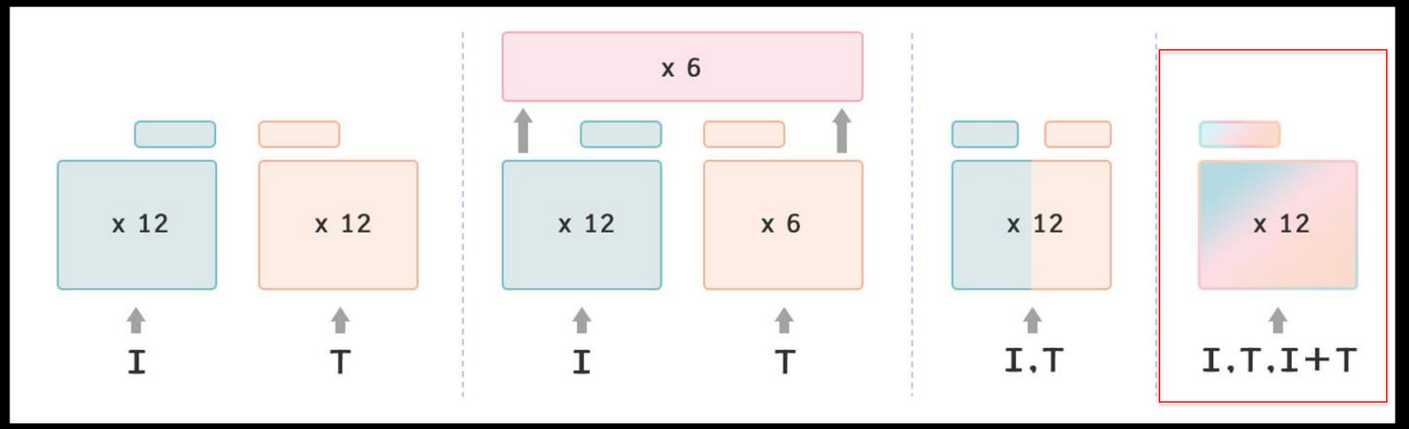
3.3 Modality-agnostic Model

NAVER
DEVIEW
2023

Goal

- 하나의 모델로 이미지와 텍스트 모두에 대해 좋은 representation 학습
- 정보 처리 과정에서 이미지와 텍스트를 서로 동일한 방식으로 활용

→ Image, text encoder가 weight share



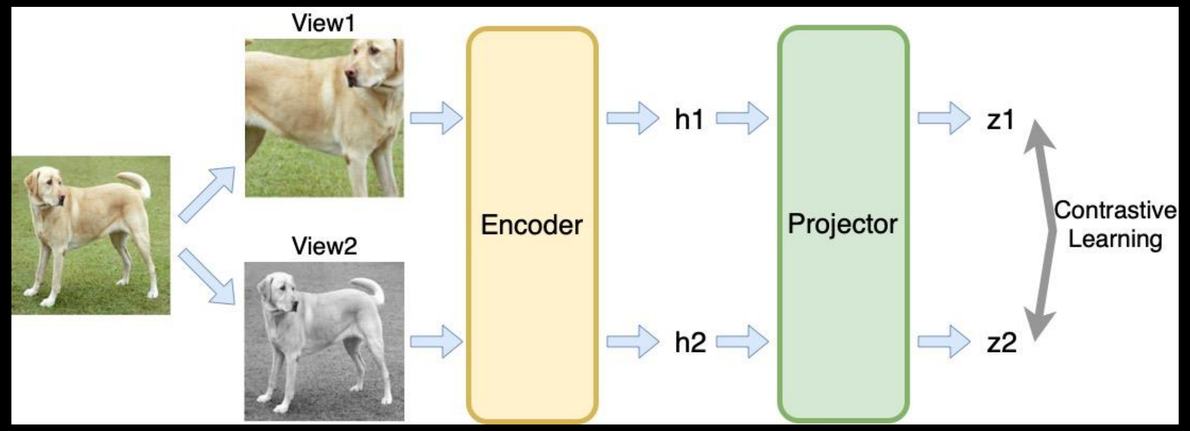
Prototypes

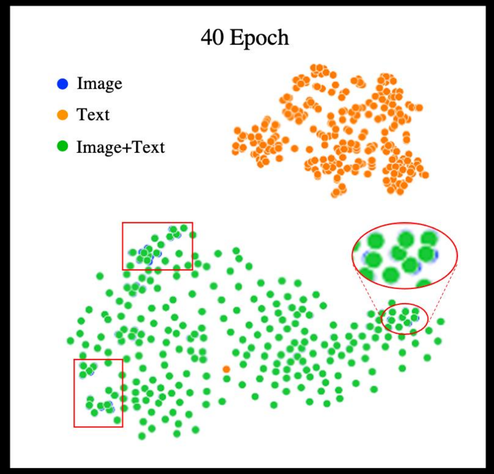
Unifying Vision-Language Representation Space with Single-tower Transformer, AAAI 2023

3.3 Modality-agnostic Model

NAVER
DEVIEW
2023

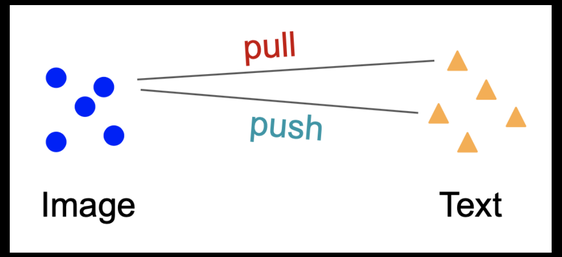
- Moco v3 의 Contrastive Self-supervised Learning framework 를 사용





Naïve Approach: Modality Gap을 극복하지 못하고 학습 실패

Image와 Text는 (특히 학습 초반에) 정보의 성격이 굉장히 이질적!



3.3 Modality-agnostic Model

NAVER DEVIEW 2023

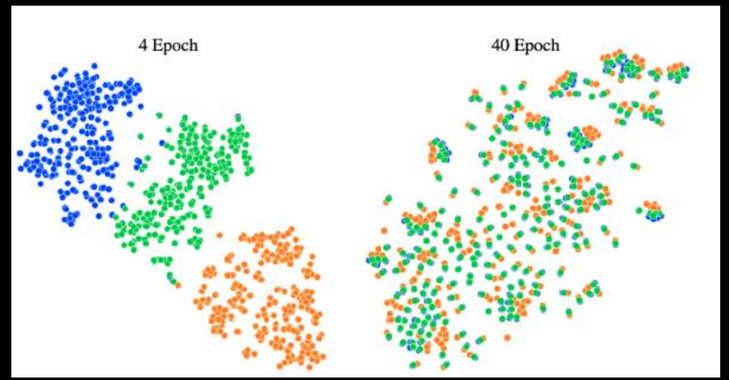
- CTR(image, text)가 modality gap 때문에 어려우니 CTR(image+text, image+text)로 학습
  student teacher → student teacher

- Vision 분야에서 널리 사용되던 Mixup을 활용 → “Cross-Modal Mixup (XMC)”

$$\mathcal{L}_{XMC} = ctr(\frac{\mathcal{F}(I)+\mathcal{F}(T)}{2}, \frac{\mathcal{F}(I)+\mathcal{F}(T)}{2}),$$

- Cross-modal Mixup으로 두 modality 간의 차이를 좁혀주면 모델은 스스로 학습 가능!
- Vision에서의 mixup이 주로 decision boundary를 smoothing하기 위한 목적이었다면,

XMC는 이질적인 두 그룹을 하나의 common ground로 projection



3.3 Modality-agnostic Model

NAVER
DEVIEW
2023

Contextual Invariance

정보 처리 과정에서 이미지와 텍스트를 서로 동일한 방식으로 활용

- F(image|text) = F(image|image) 가 된다면 modality-agnostic reasoning에 가까워질 수 있음.
- 이 때 F(X|Y)는 Y를 context (key, value)로 했을 때 X (query)의 최종 representation ([cls] 토큰 값)

- 이 직관을 앞서 등장한 XMC와 결합하여, $\mathcal{L}_{CIC} = ctr(\frac{\mathcal{F}(I|T) + \mathcal{F}(T|I)}{2}, \frac{\mathcal{F}(I|I) + \mathcal{F}(T|T)}{2}),$

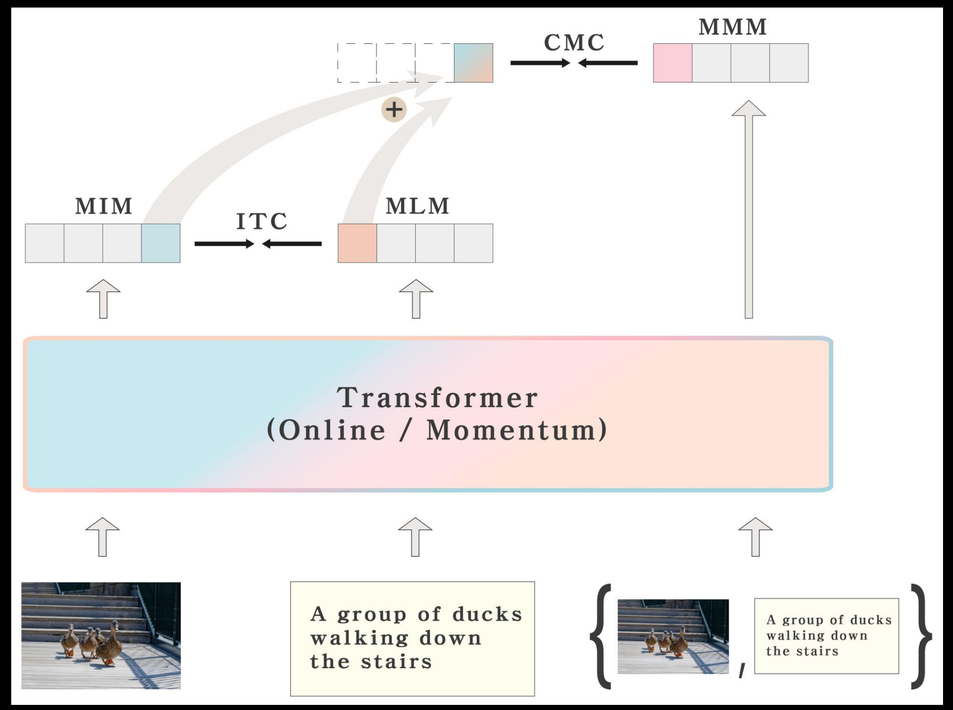
3.3 Modality-agnostic Model

NAVER DEVIEW 2023

Contextual Mixup Contrast

- CIC의 objective를 보존하되 보다 단순하고 일반화된 형태의 CMC 를 최종 제안

$$\mathcal{L}_{CMC} = ctr(\mathcal{F}(I,T|I,T), \frac{\mathcal{F}(I|I)+\mathcal{F}(T|T)}{2})$$

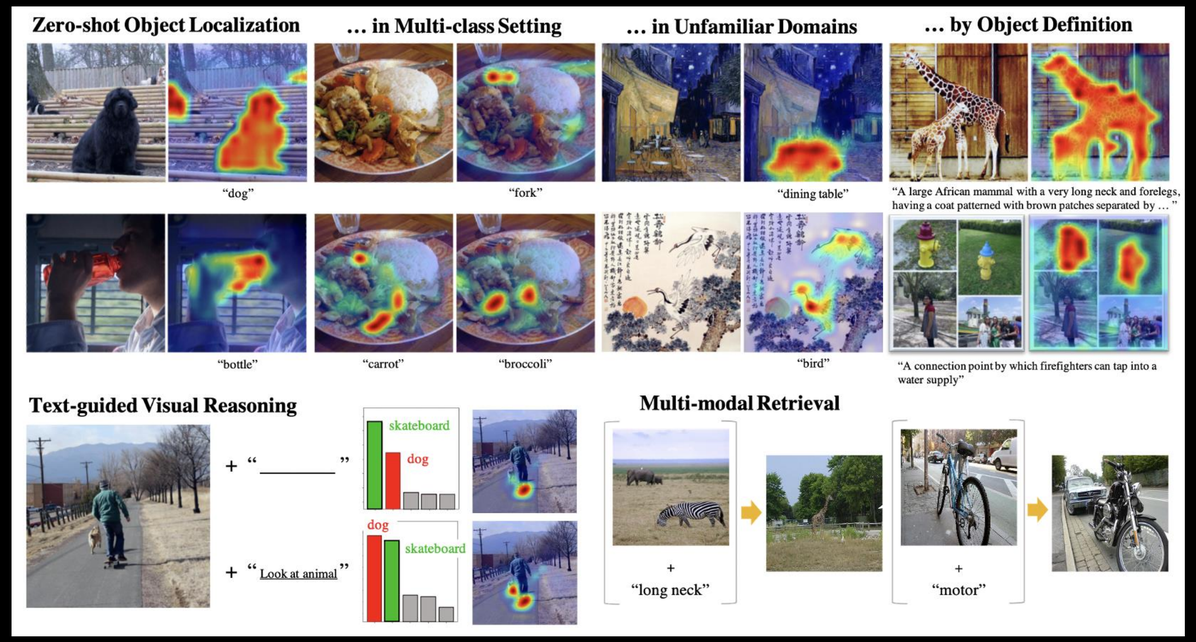


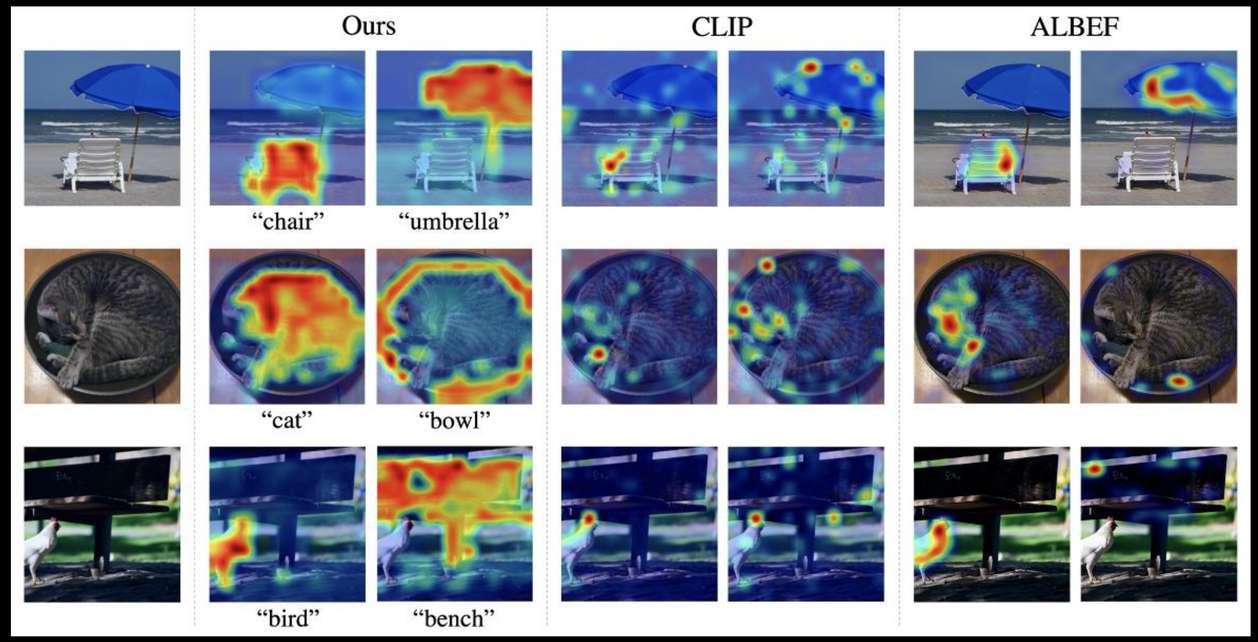
Method	Imagenet	MS-COCO			
	Top-1 Acc.	TR@1	TR@5	IR@1	IR@5
CLIP	17.1	15.0	34.8	10.9	26.7
SLIP	23.0	21.7	45.1	15.6	35.2
ITC	1.6	0.8	2.5	0.7	2.2
ITC (two heads)	17.5	10.4	26.8	10.7	26.4
ITC + XMC	22.1	25.2	48.1	15.2	33.6
ITC + XMC + CIC	22.9	25.4	48.1	16.3	35.5
ITC + CMC (OneR)	**23.7**	**25.5**	**48.2**	**16.9**	**36.9**

3.3 Modality-agnostic Model

NAVER
DEVIEW
2023

성능 벤치마킹





Text Embedding과 Image Patch들 사이의 similarity 를 시각화하는 것 만으로 Object localization 수준의 정교한 특정이 가능

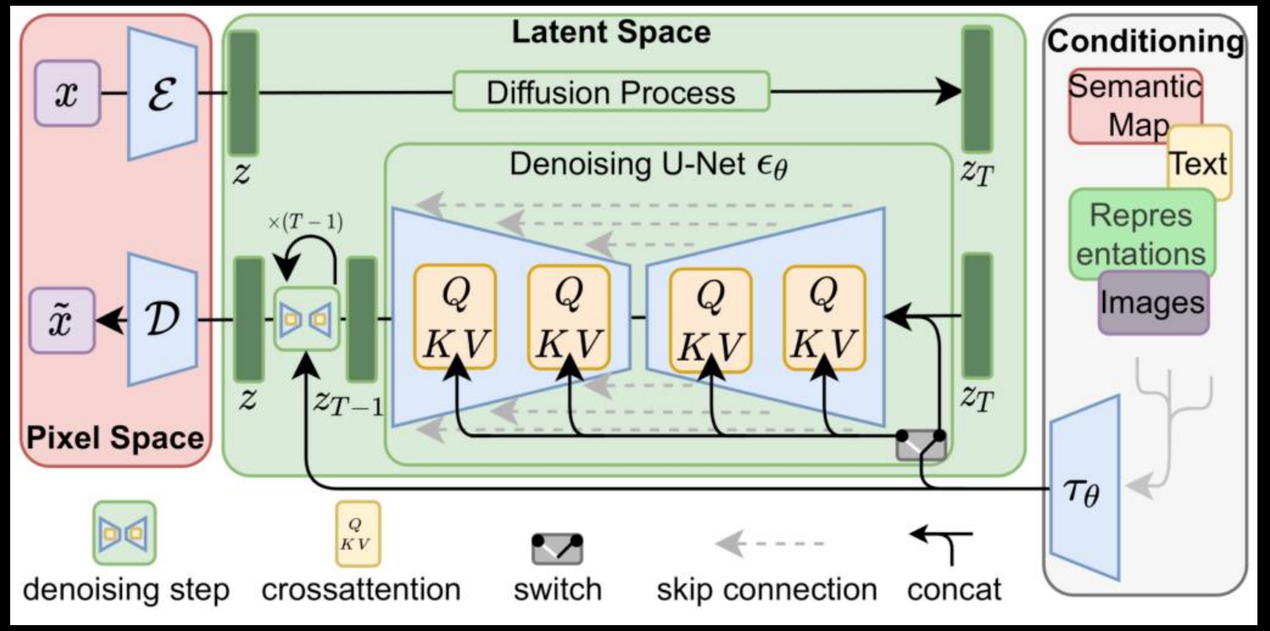
CLIP이나 ALBEF의 경우 image, text를 같은 공간에 embed 하지 않으므로 similarity operation은 불가능하고, Grad-CAM을 이용하여 시각화

4. Korean Text-to-Image Generation

4.1 Stable diffusion

NAVER
DEVIEW
2023

- Auto encoder를 활용해 pixel space가 아닌 latent space에서 diffusion process를 진행
- Text를 활용하는 경우, text encoder를 통해 나온 embedding을 denoising 과정에서 attention을 계산할 때 넣어줌으로써 guidance를 제공



High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022

4.2 Text encoder training

NAVER
DEVIEW
2023

기존 stable diffusion model의 한계점

- Text encoder로 쓰고 있는 CLIP 모델이 영어 데이터로 학습됨
- 한글 prompt를 사용해 생성하면 엉뚱하고 기괴한 이미지 생성

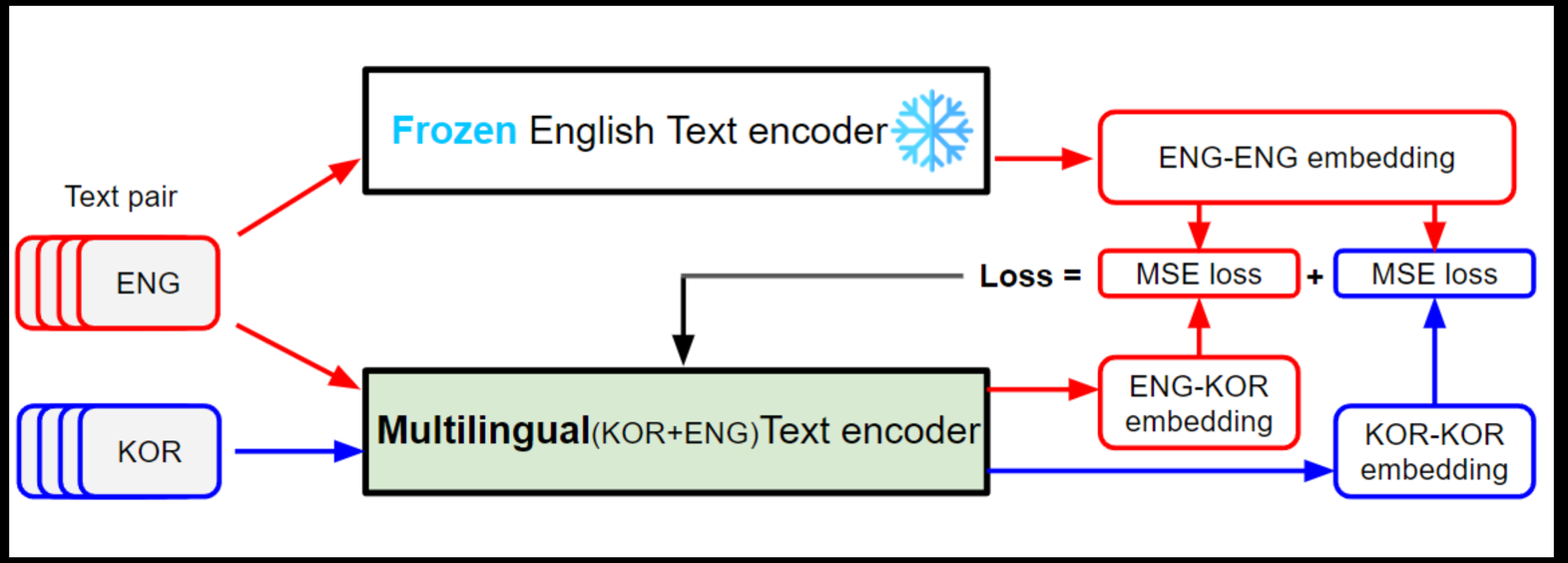
Prompt: “강아지 사진 "

4.2 Text encoder training

NAVER
DEVIEW
2023

학습과정

- Knowledge distillation을 활용해 기존 English latent space에 Korean embedding도 활용할 수 있도록 유도
- 학습 후 English, Korean 모두 같은 latent space상에서 표현 가능

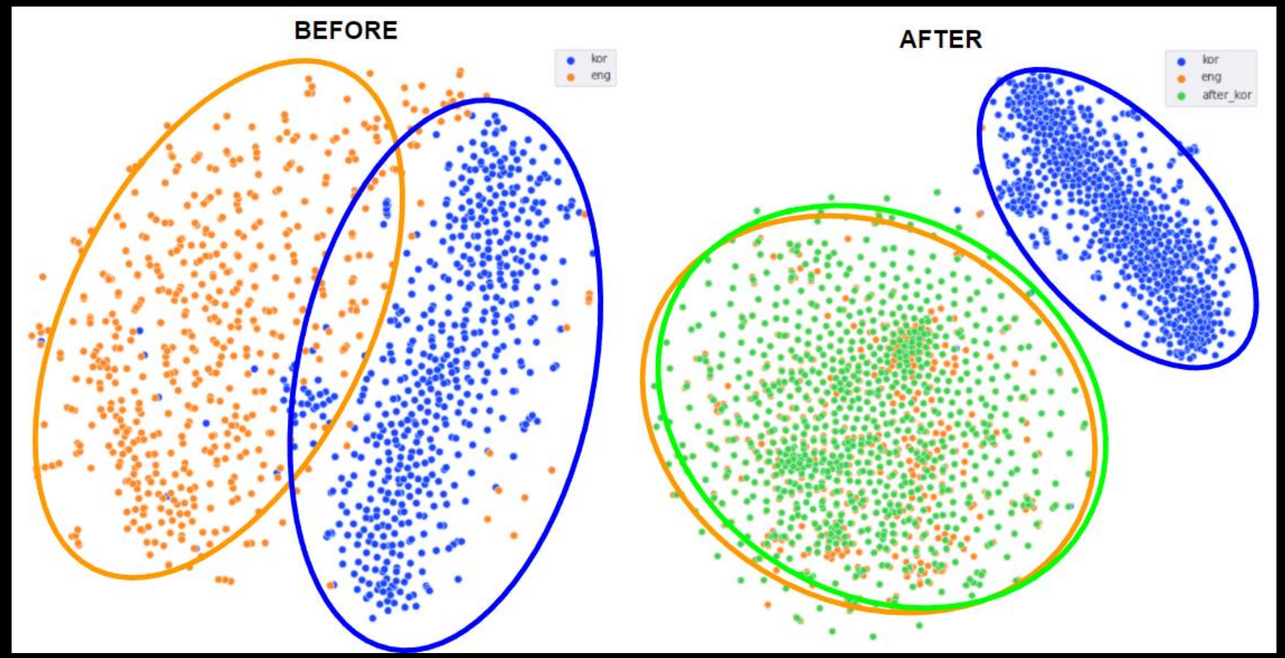


4.2 Text encoder training

NAVER
DEVIEW
2023

TSNE 시각화

- Original model(CLIP L/14) vs Multilingual Text encoder간 text embedding 비교



다른 언어 사이의 gap이 명확히 존재

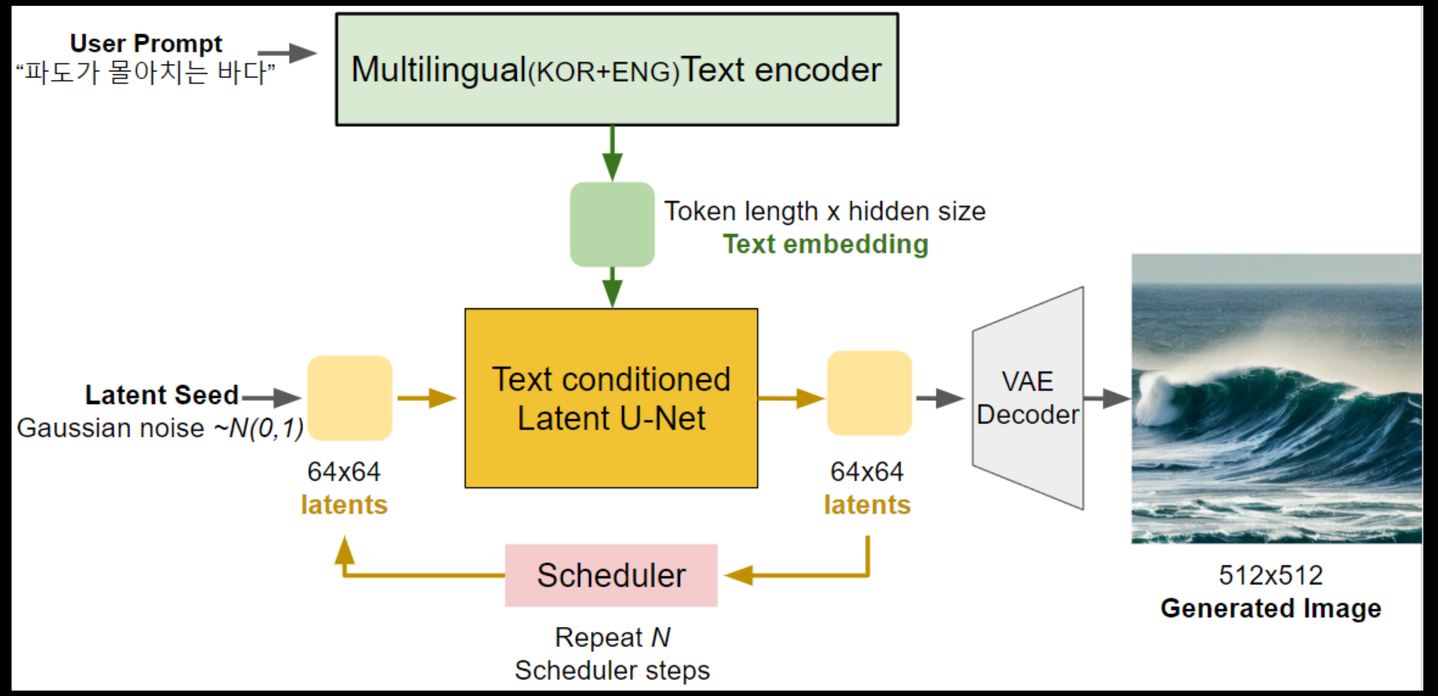
동일한 데이터에 대해 English embedding과
Korean embedding이 같은 latent space에 위치

4.3 Text2Image Inference

NAVER
DEVIEW
2023

Multilingual text encoder에서 embedding을 추출한 뒤 denoising 과정에서 cross-attention

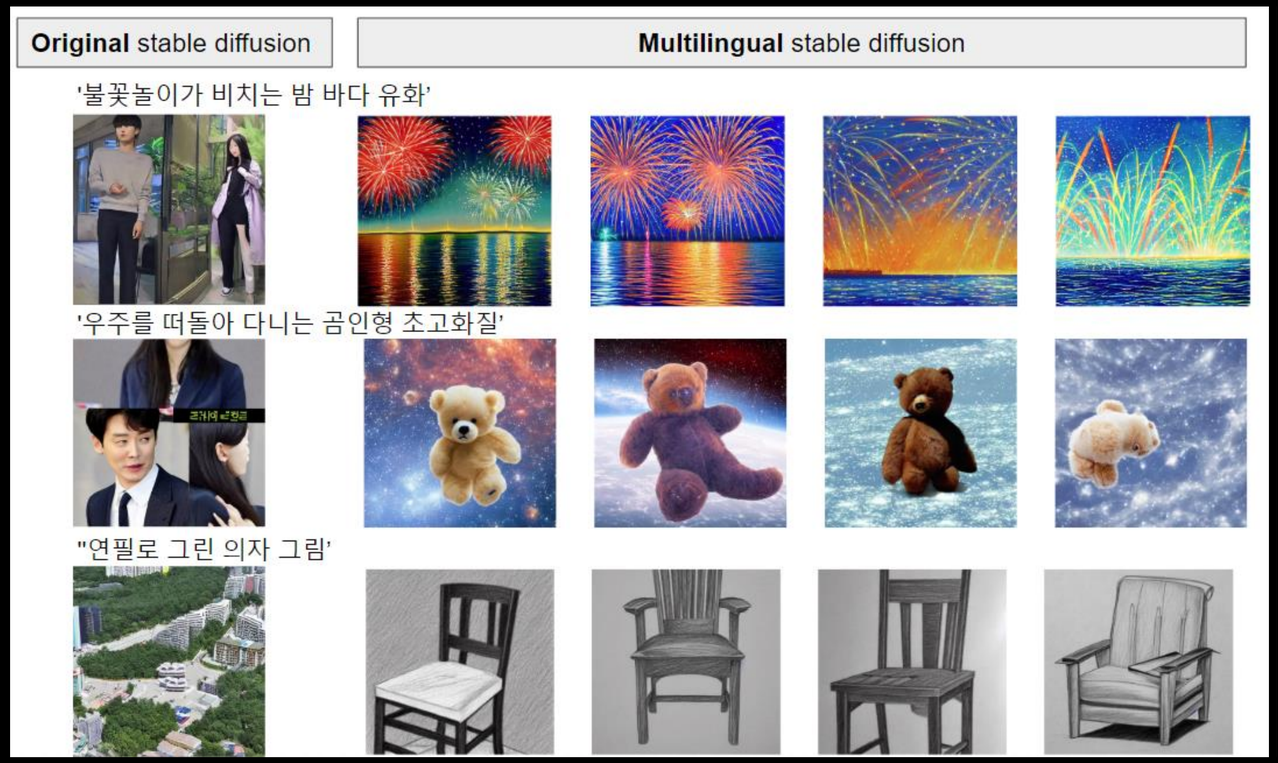
- 일반적으로 50 step 정도 반복, 1장당 2-3초의 inference time 소요 (A100 기준)

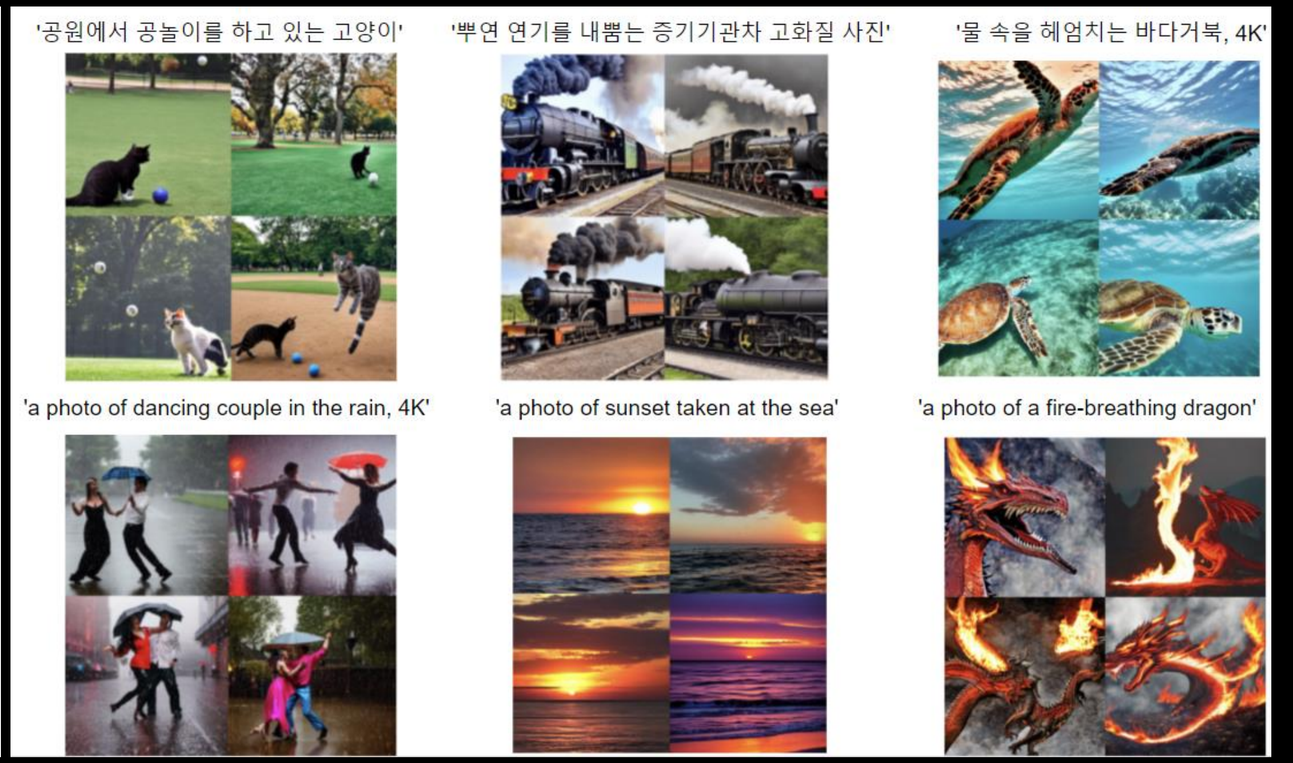


4.3 Text2Image Inference

NAVER
DEVIEW
2023

생성 예시



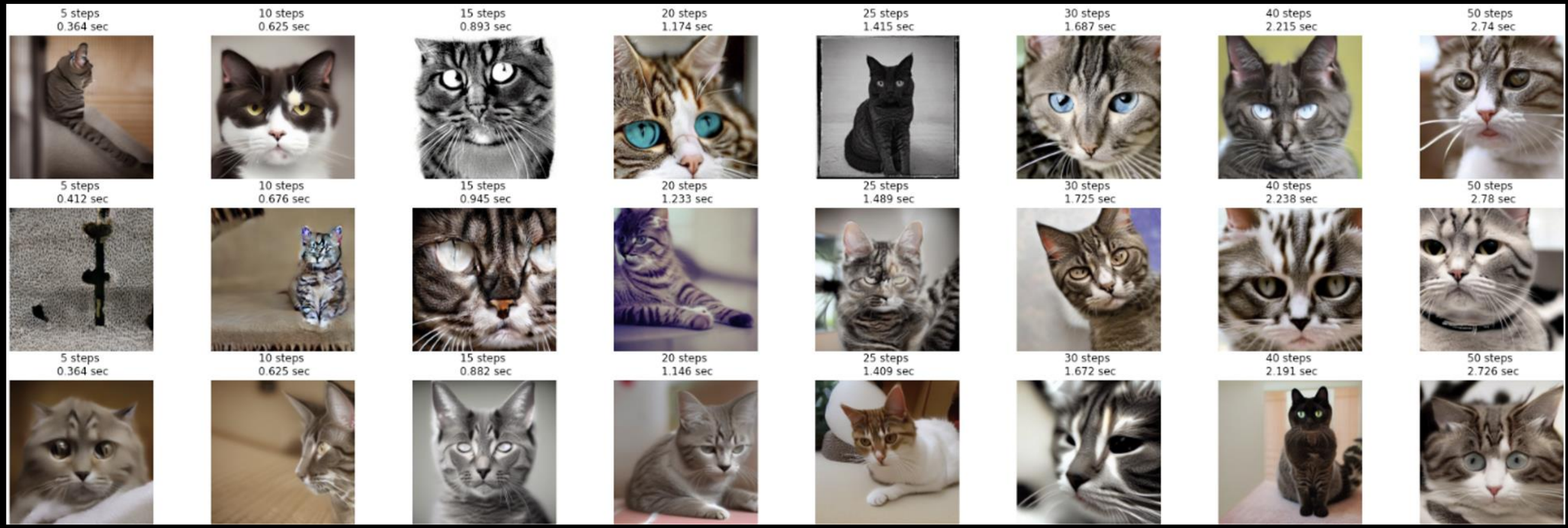


4.3 Text2Image Inference

NAVER
DEVIEW
2023

A100 기준 생성 속도(step 수)별 품질

- Prompt: ‘고양이 사진’



4.3 Text2Image Inference

NAVER
DEVIEW
2023

A100 기준 생성 속도(step 수)별 품질

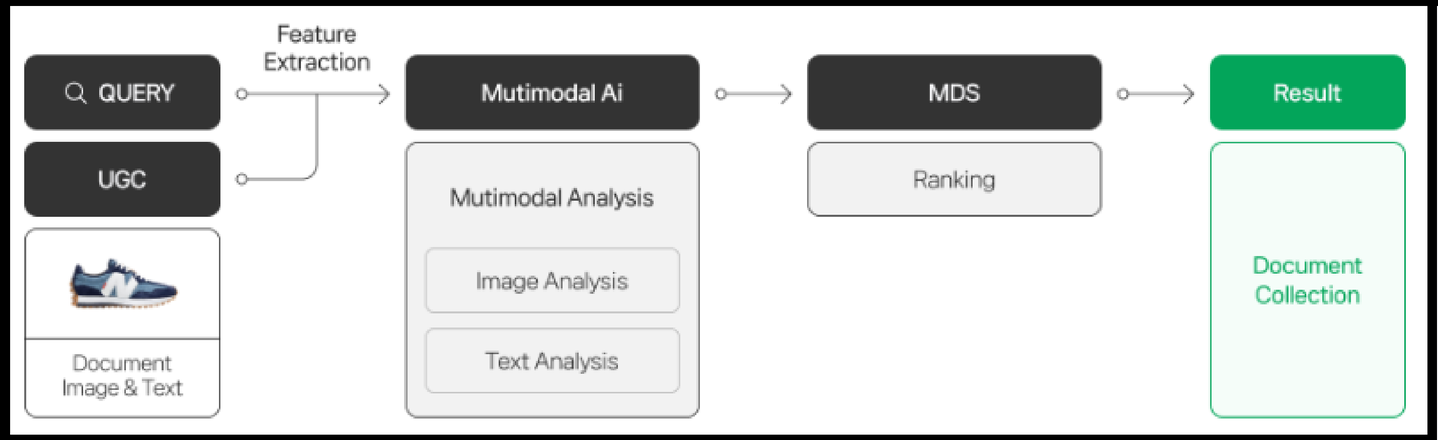
- Prompt: ‘파란색 드레스’

5. Multimodal Document Search (MDS) 서비스 적용

5.1 Multimodal Document Search (MDS)

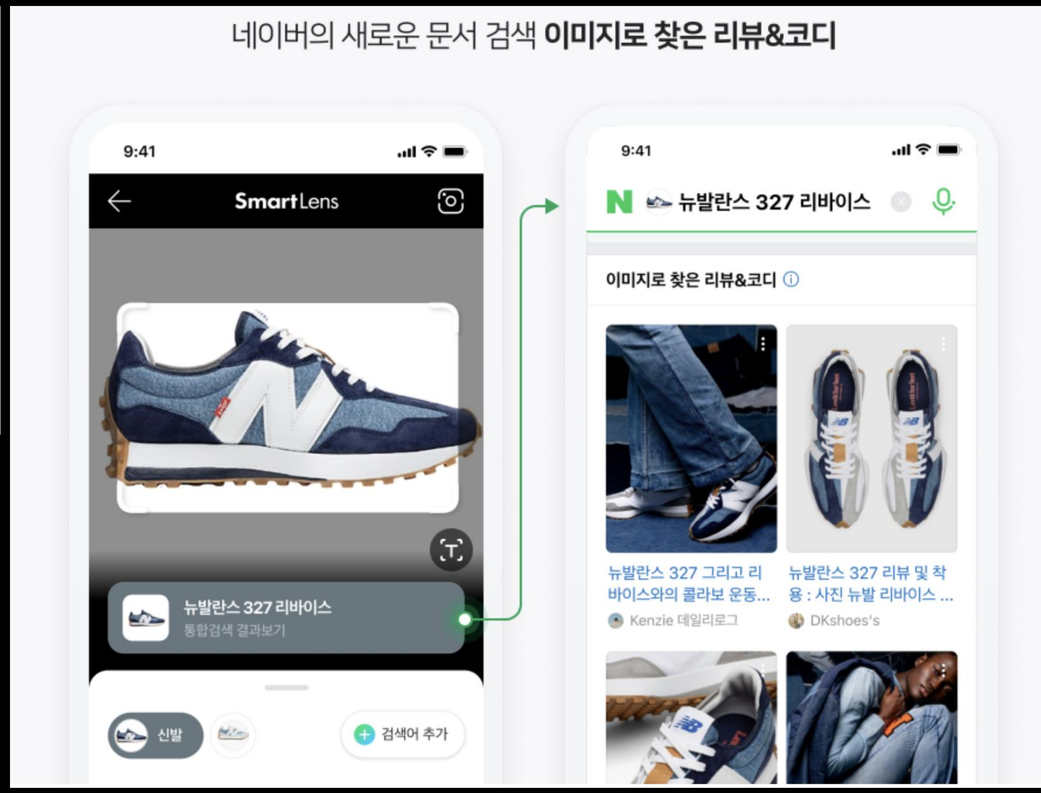
NAVER
DEVIEW
2023

네이버의 멀티모달 문서 검색 서비스



- 이미지 vectorization에 VLM 활용
- 멀티모달 질의에 대한 문서 검색 및 랭킹 → Text Neural Matching & Image Neural Matching

예시



5.1 Multimodal Document Search (MDS)

NAVER
DEVIEW
2023

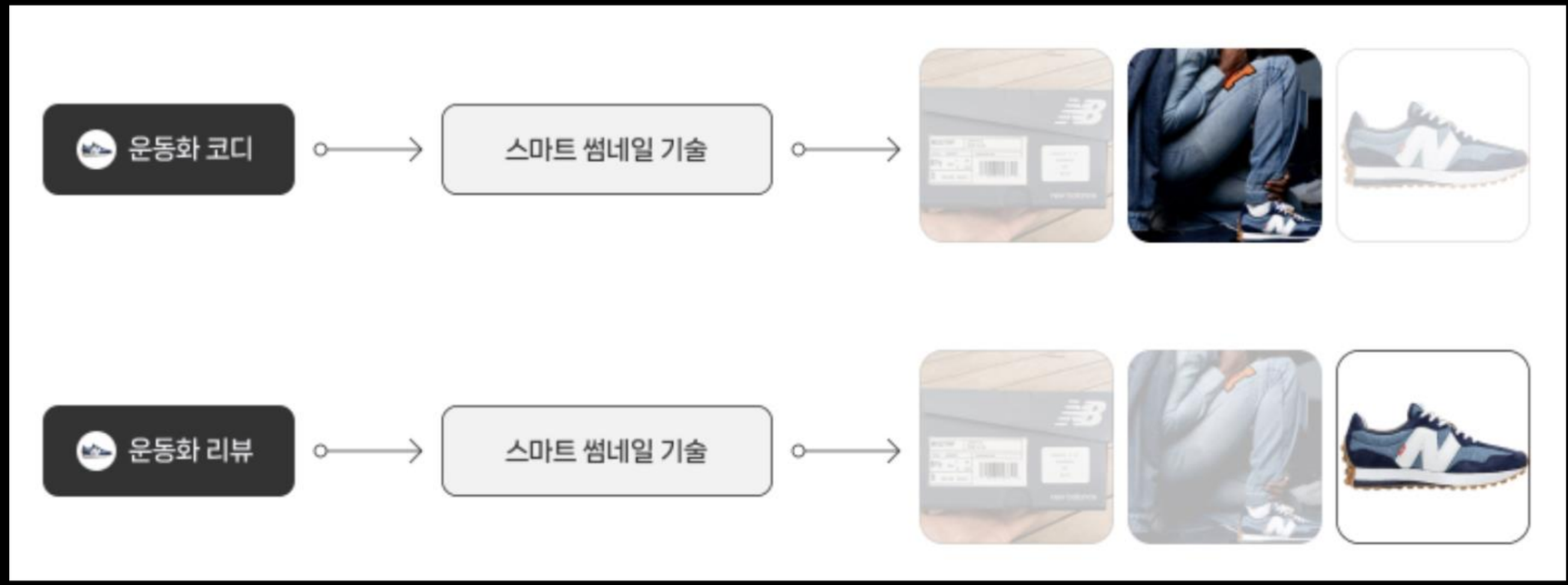
멀티모달 관련 요소 기술들

1. 스마트 썸네일
2. 멀티모달 기반 패션 상품 검색
3. 의류 검색을 위한 색상/패턴 분류기

5.2 스마트 썸네일

NAVER
DEVIEW
2023

쿼리-이미지 점수로 문서 내 이미지들을 ranking 및 selection



5.2 스마트 썸네일

NAVER
DEVIEW
2023

Image score 평가 기준

5 : 쿼리와 연관성이 있으며 (주제, 의도 연관성 있음) 스니펫 품질에 도움을 줌

4 : 쿼리와 연관성이 있으나 (주제, 의도 연관성 있음) 품질상 도움을 별로 주지 않음

3 : 쿼리와 연관성이 약함 (주제 정도만 연관성 있음)

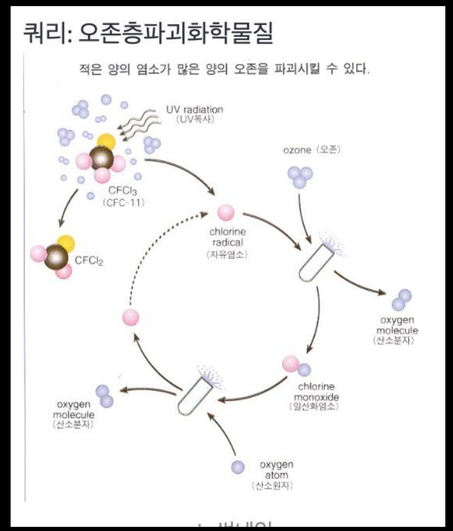
2 (저품질) : 쿼리와 아예 연관이 없는 이미지

1 (저품질) : 쿼리와 아예 연관이 없고 서비스 나가면 위험한 이미지

2 점

3 점

4 점



5 점

5.2 스마트 썸네일

NAVER
DEVIEW
2023

성능 벤치마킹

- image-text similarity score 일정 이하는 필터링

	ASIS	Naver_VLM_base	Naver_VLM_large
Coverage	60	49 (81.7%)	44 (73.3%)
Image_score > 2	47	45	42
Image_score > 3	13 (21.7%)	4 (8.2%)	2 (4.5%)
필터 후 Image_score 평균	3.133	3.327	3.455

5.3 멀티모달 기반 패션 상품 검색

NAVER
DEVIEW
2023

패션 상품 검색 예시

원본 쿼리 이미지

Fashion Object Detector

Crop 이미지

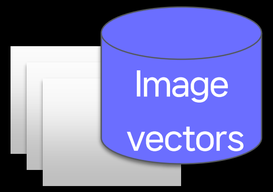
Image Feature

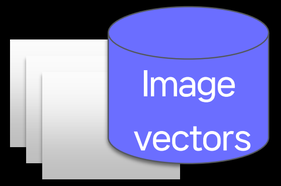
Naver_VLM1

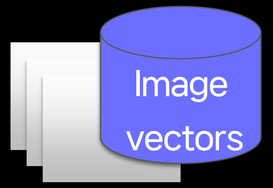
Naver_VLM2

Naver_VLM3

Image Retrieval







검색된 Top-k 상품

나이키 에어조던 1 ko 시카고

나이키 X NBA 덩크 로우 레트로 EMB 시카고

나이키 덩크 하이 범고래 블랙

...

Image-Text Reranking

Top1 상품

나이키 X NBA 덩크 로우 레트로 EMB 시카고

5.3 멀티모달 기반 패션 상품 검색

NAVER
DEVIEW
2023

의류 검색 예시

- 신발의 시리즈명 같이 명확한 상품 이름이 없는 문제

Top1 상품

보헤미안 케이프세트 원피스

VLM 기반 색상 및 패턴을 분류, 문서검색용 메타데이터로 함께 태깅

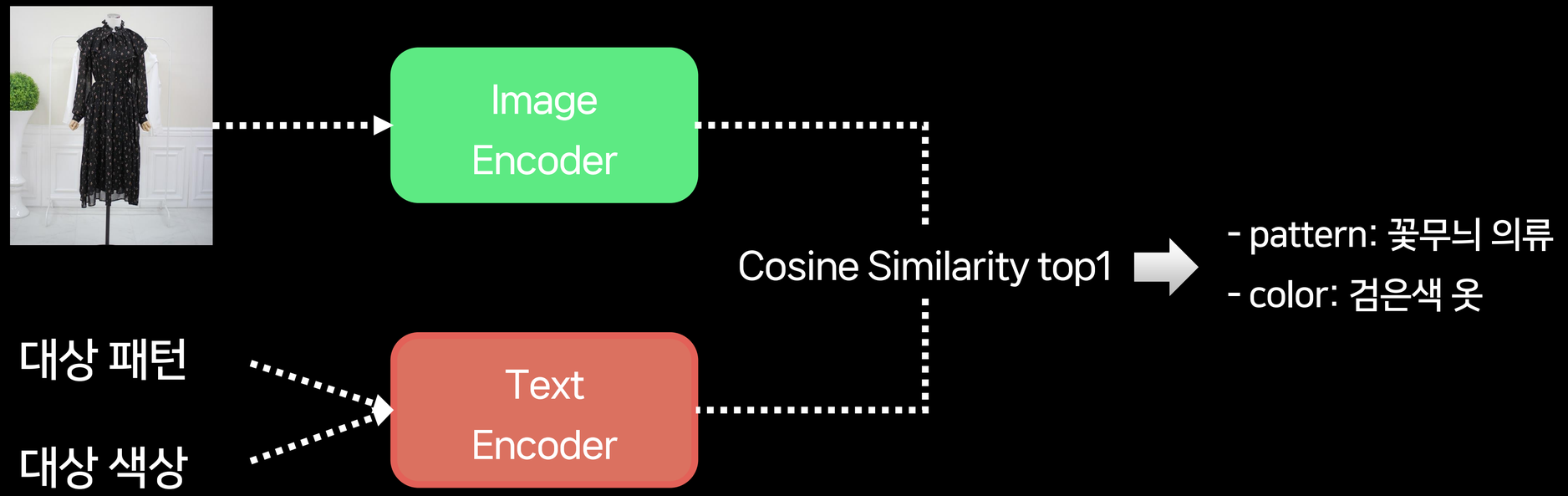
- pattern: 꽃무늬 의류
- color: 검은색 옷

5.3 멀티모달 기반 패션 상품 검색

NAVER
DEVIEW
2023

의류 검색을 위한 색상 및 패턴 분류기

- 대상 색상 = ["네이비색 옷", "보라색 옷", "빨간색 옷", "베이지색 옷", "검은색 옷", "녹색 옷", "파란색 옷", "흰색 옷", …]
- 대상 패턴 = ["민무늬 의류", "꽃무늬 의류", "도트무늬 의류", "별무늬 의류", "하트무늬 의류", "줄무늬 의류", "체크무늬 의류", …]



5.3 멀티모달 기반 패션 상품 검색

NAVER
DEVIEW
2023

의류 색상 및 패턴 분류기 개선

- 일부 Label들의 학습 샘플이 부족한 data imbalance 문제를 이미지 생성 기반 Data augmentation으로 해결

5.4 Future Works

NAVER
DEVIEW
2023

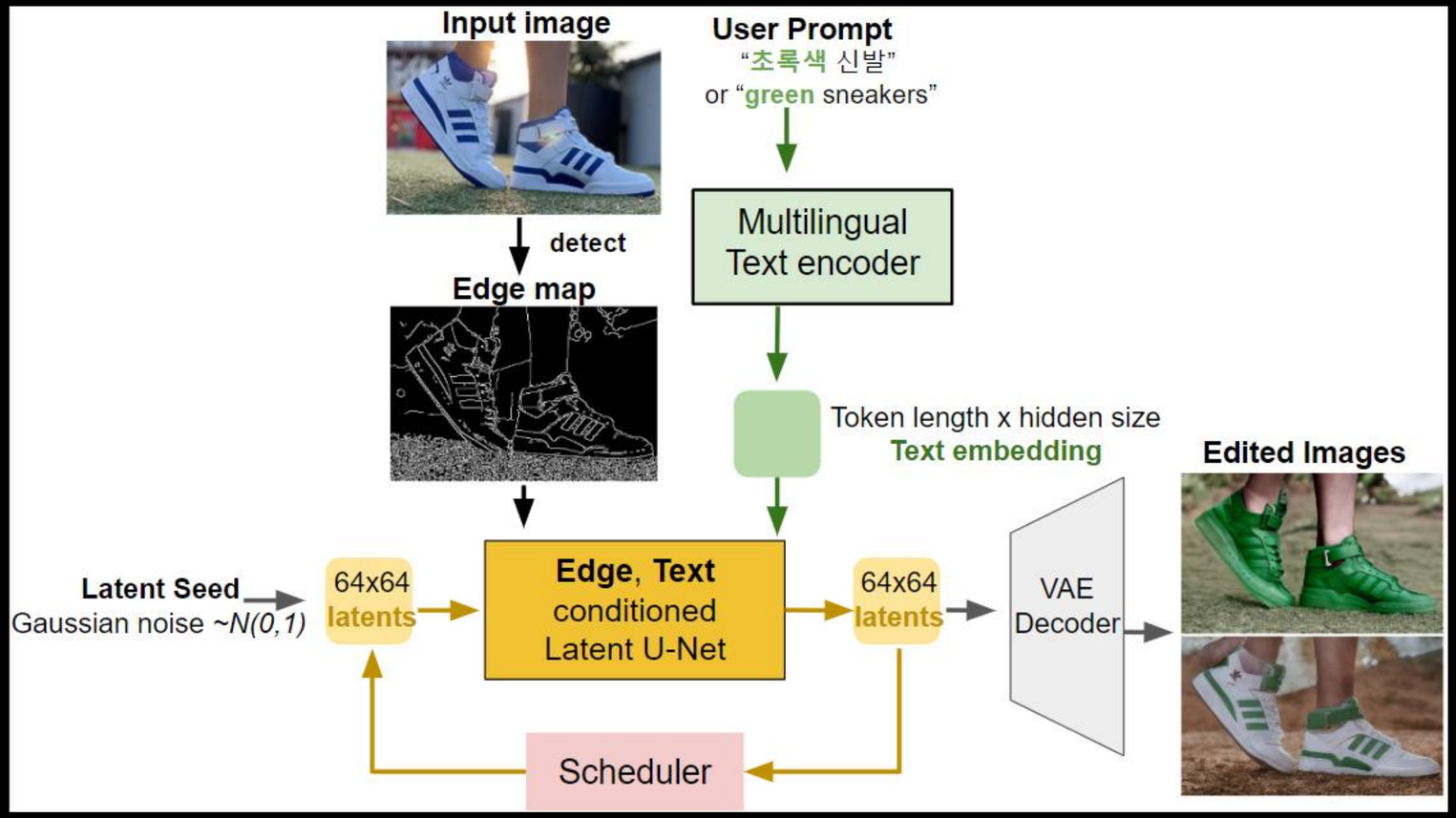
- 모델은 더 크게, 데이터는 더 많이
- 더 다양한 태스크에 적용 (Visual Q&A/대화, ...)
- MDS 대상 카테고리 및 주제 확장

...

5.4 Future Works

NAVER
DEVIEW
2023

Fashion image to image translation



✚ 멀티모달 기반 패션 상품 검색

Adding Conditional Control to Text-to-Image Diffusion Models, 2023

5.4 Future Works

NAVER
DEVIEW
2023

MDS 명소 검색

명소 image-text DB

Input Image

VLM

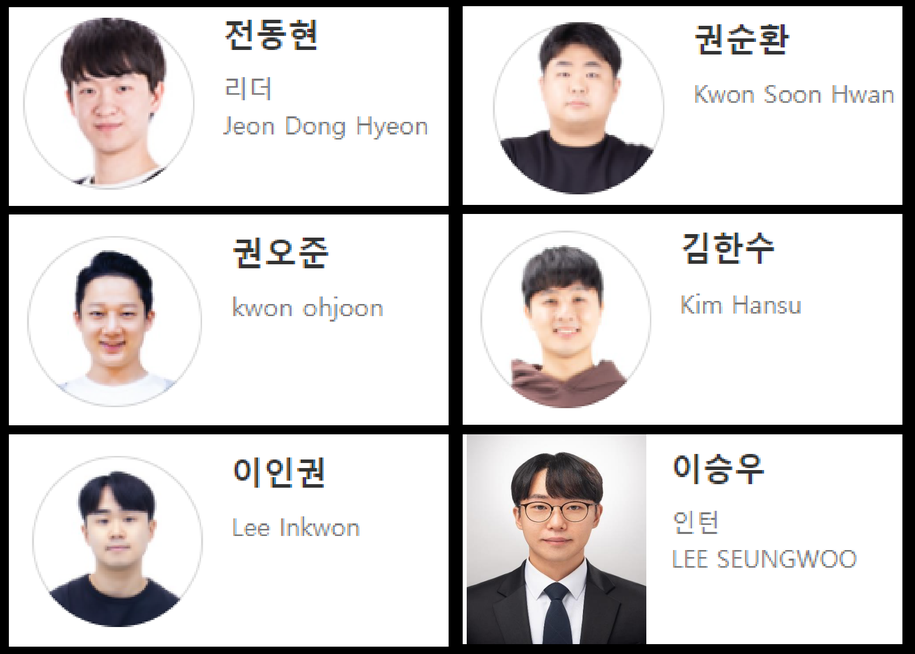
Top1: 마리나베이 샌즈

Top2: 마리나베이, 싱가포르

...

MDS

Thank You



donghyeon.jeon@navercorp.com