DALLE Couture: Text to Fashion

#Text-to-Image #KoDALLE #KoCLIP

NLP-15
(HappyFace팀)

boostcamp ai tech

Executive Summary

DALLE Couture: Text to Fashion

Why

- Prototyping and redesigning clothes in similar manners are pain points in fashion industry.
- Yielding multiple candidates for the similar designs is time consuming.
- Loop of designing, 3D modeling, hiring fit models and analyzing is costly.

What

- **Trained and utilized VQGAN and KoDALLE for Korean text to fashion image(250x250) generation.**
- **Reranked images generated with KoDALLE by restructuring and training KoCLIP.**

How

[Phase 1: Image Generation]

- Compared VAE, VQVAE and VQGAN for high resolution image generation.
- Trained VQGAN model from scratch on K-Fashion Image Dataset of AIHub.

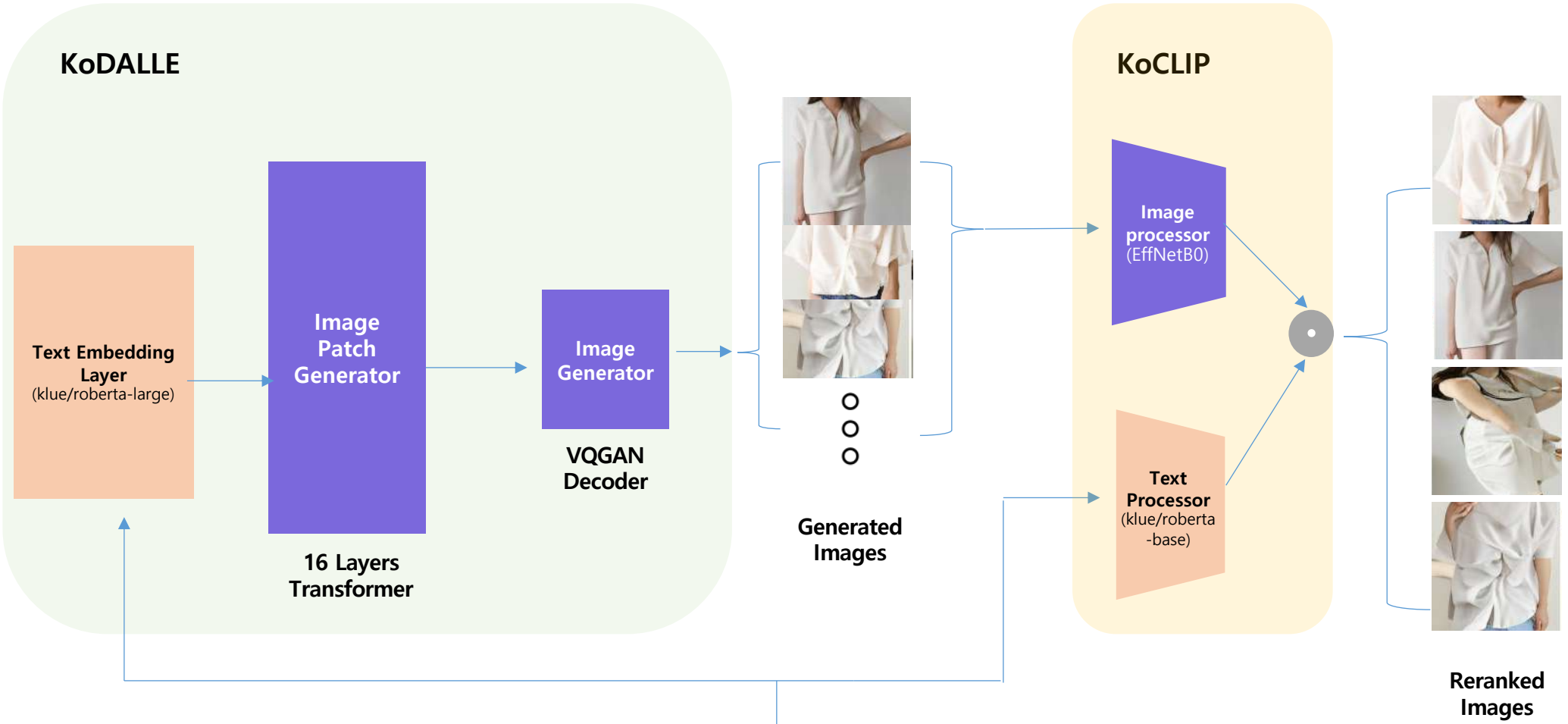
[Phase 2: Text to Image]

- Compared AttentionGAN and DALLE for multimodal text to image models.
- Adapted DALLE model with Korean text embedding layer for robust inference: KoDALLE.
- Trained KoDALLE Model from scratch with K-Fashion Image Dataset with text pairs.

[Phase 3: Rerank and Deploy]

- Constructed EfficientNetB0 + klue/roberta-base for algorithmic reranking: KoCLIP.
- Deployed demo of KoDALLE model using streamlit on Huggingface Hub.

Executive Summary



	OpenAI's DALLE	KoDALLE of HappyFace
#Params	12 Billion	428 Million
#Layers	64 Layers	16 Layers
Computing Resource	1024 x V100 16GB	1 x V100 32GB
Text Encoder	16384 Vocab x 512 Dim BPE	32000 Vocab x 1024 Dim klue/roberta-large
Image Encoder	VQVAE	VQGAN
Optimizer	AdamW	AdamW
Learning Rate	4.5e-5	3.0e-5
Weight Decay	4.5e-3	3.0e-3
LR Scheduler	ReduceLROnPlateau	-

Table of Contents

Background

Project Background

For fashion industry, it is required to
Design, Build, Fit to models(or mannequins) and Reproduce with different conditions.

Project Background

Before 3D modeling one particular design(e.g. CLO 3D),
Prototyping multiple candidate designs with text conditions will save time for fashion designers.

Result

Result

하의에서 색상은 스카이블루이다. 상의에서 기장은 롱이다. 색상은 화이트이다. 카테고리는 블라우스이다. 디테일에는 셔링이다. 소매기장은 반팔이다. 소재에는 실크이다. 프린트에는 무지이다. 넥라인은 브이넥이다. 핏은 노멀

상의에서 기장은 노멀이다. 상의에서 색상은 화이트이다. 상의에서 서브색상은 블랙이다. 상의에서 카테고리는 티셔츠이다. 상의에서 소매기장은 반팔이다. 상의에서 소재에는 저지이다. 상의에서 프린트에는 레터링이다. 상의에서 넥라인은 라운드넥이다. 상의에서 핏은 루즈이다.

Input Text Sequence and Generated Images(250 x 250)

Dataset


```
"렉트좌표": {
        "아우터": [{ "X좌표": 202.5, "Y좌표": 68.5, "가로": 416, "세로": 476 }],
        "하의": [{}],
        "원피스": [{ "X좌표": 265.5, "Y좌표": 73.5, "가로": 252, "세로": 525 }],
        "상의": [{}]
      },
"라벨링": {
        "스타일": [{ "스타일": "로맨틱" }],
        "아우터": [
          {
            "기장": "노말",
            "색상": "화이트",
            "서브색상": "민트",
            "카테고리": "재킷",
            "소매기장": "긴팔",
            "소재": ["트위드"],
            "프린트": ["체크"],
            "넥라인": "라운드넥",
            "핏": "노멀"
          }
        ],
        "하의": [{}],
        "원피스": [
          {
            "카테고리": "드레스",
            "디테일": ["단추"],
            "소재": ["우븐"],
            "프린트": ["무지"],
            "넥라인": "브이넥",
            "핏": "노멀"
          }
        ],
        "상의": [{}]
      }
    },
```


K-Fashion Dataset: 1 million images paired with labels of coordinates and classes.
To resolve class imbalance, 0.77 million images and text pairs were chosen for training and 0.05 million images were chosen for test validation.

250 x 250

Cropped images of objects based on coordinates and resized to 250 x 250.

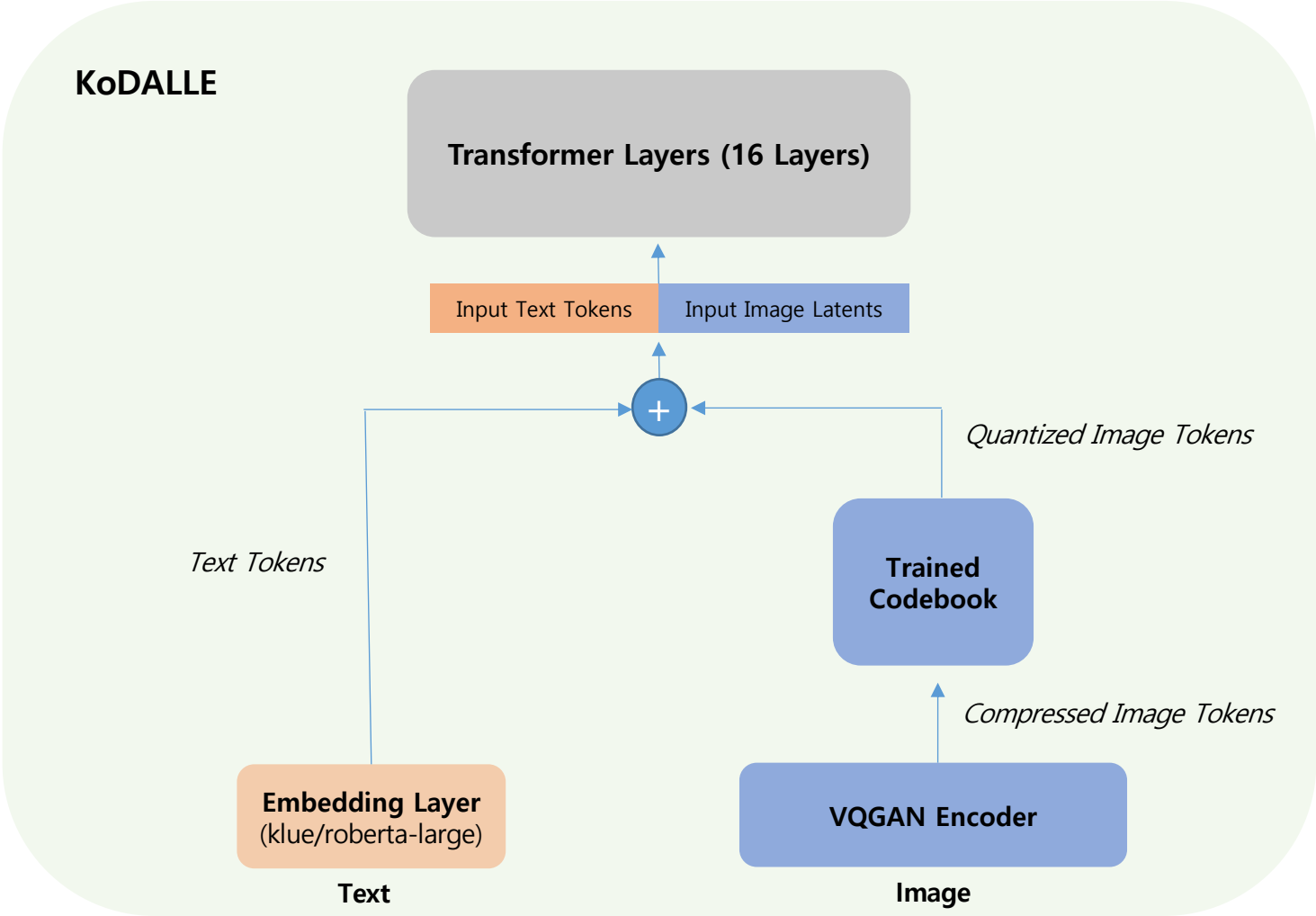
```
"렉트좌표": {
        "아우터": [{ "X좌표": 202.5, "Y좌표": 68.5, "가로": 416, "세로": 476 }],
        "하의": [{}],
        "원피스": [{ "X좌표": 265.5, "Y좌표": 73.5, "가로": 252, "세로": 525 }],
        "상의": [{}]
      },
"라벨링": {
        "스타일": [{ "스타일": "로맨틱" }],
        "아우터": [
          {
            "기장": "노말",
            "색상": "화이트",
            "서브색상": "민트",
            "카테고리": "재킷",
            "소매기장": "긴팔",
            "소재": ["트위드"],
            "프린트": ["체크"],
            "넥라인": "라운드넥",
            "핏": "노멀"
          }
        ],
        "하의": [{}],
        "원피스": [
          {
            "카테고리": "드레스",
            "디테일": ["단추"],
            "소재": ["우븐"],
            "프린트": ["무지"],
            "넥라인": "브이넥",
            "핏": "노멀"
          }
        ],
        "상의": [{}]
      }
    },
```


아우터에서 기장은 노말이다. 아우터에서 색상은 화이트이다. 아우터에서 서브색상은 민트이다. 아우터에서 카테고리는 재킷이다. 아우터에서 소매기장은 긴팔이다. 아우터에서 소재에는 트위드이다. 아우터에서 프린트에는 체크이다. 아우터에서 넥라인은 라운드넥이다. 아우터에서 핏은 노멀이다.

원피스에서 카테고리는 드레스이다. 원피스에서 디테일은 단추이다. 원피스에서 소재는 우븐이다. 원피스에서 프린트는 무지이다. 원피스에서 넥라인은 브이넥이다. 원피스에서 핏은 노멀이다.

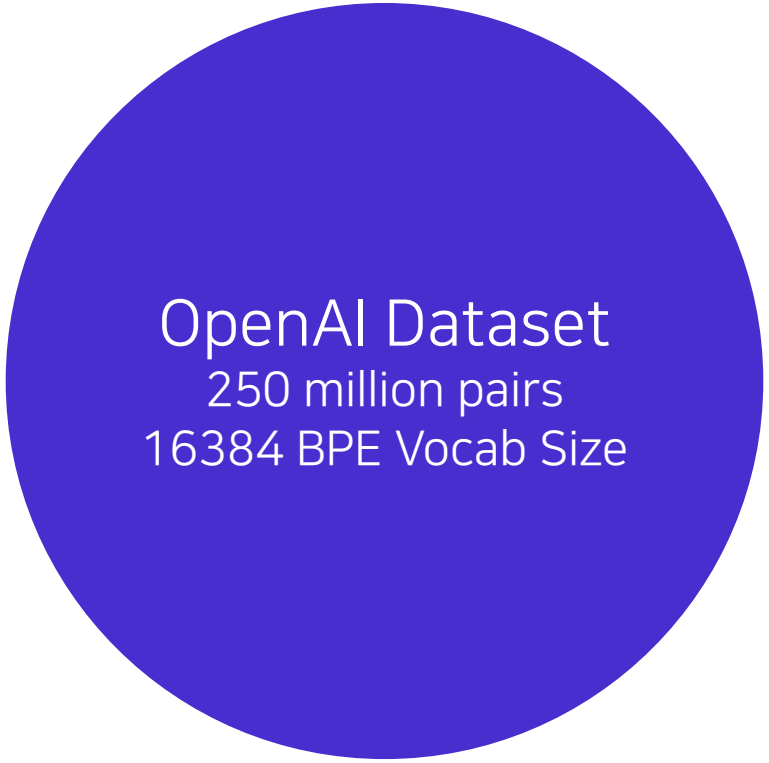
Converted JSON labels into Korean sentence labels.
Input image tokens are 256 whereas text tokens range from 64 to 128.

KoDALLE



Text Tokens and Quantized Image Tokens pass throughs 16 layers of Transformers.

KoDALLE



Training DALLE model from scratch demands large size paired dataset of images and captions.

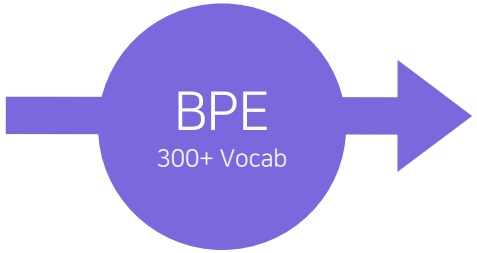
KoDALLE

K-Fashion Dataset Vocab

스카이블루
셔링
오버핏
미니

Fashion Vocab

여친룩
OOTD
쿨톤
고급스러운



Tokenized K-Fashion Vocab

스카이, ##블루
셔링
오버, ##핏
미니

Tokenized Fashion Vocab

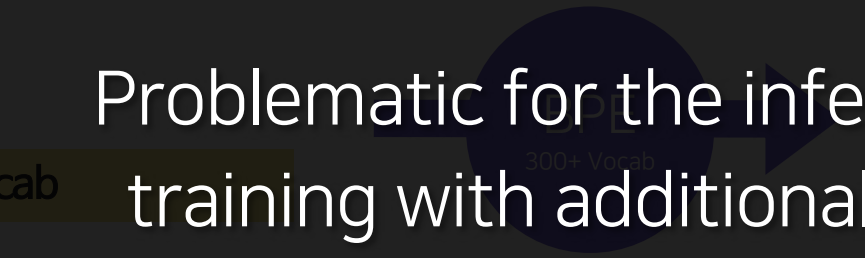
[UNK], ##룩
[UNK], [UNK], [UNK], [UNK]
[UNK], [UNK]
[UNK], ##스러운

If the training dataset isn't large enough or is limited to specific domains, number of vocabularies in the trained DALLE model are insufficient.

KoDALLE

K-Fashion Dataset Vocab	Tokenized K-Fashion Vocab
스카이블루	스카이, ##블루
셔링	셔링
오버핏	오버, ##핏
미니	미니

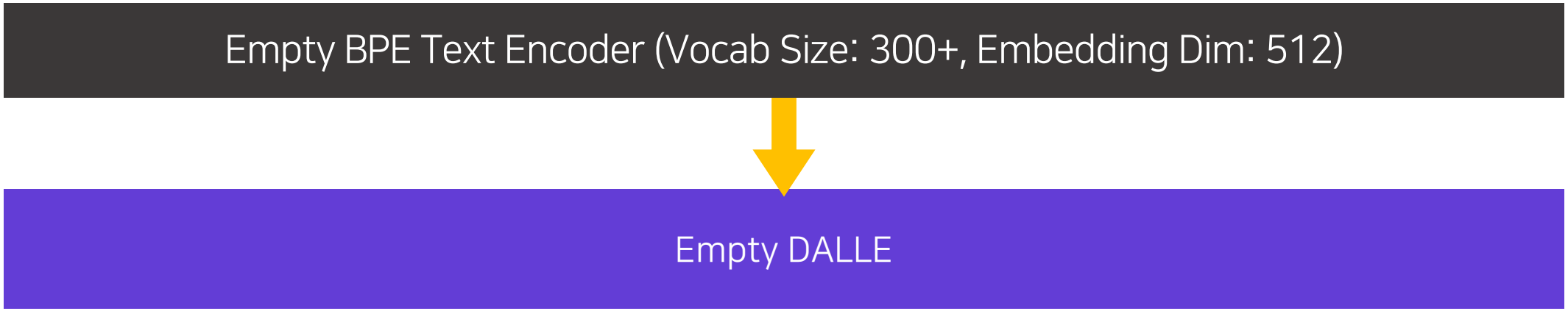
Fashion Vocab	Tokenized Fashion Vocab
여친룩	[UNK], ##룩
OOTD	[UNK], [UNK], [UNK], [UNK]
쿨톤	[UNK], [UNK]
고급스러운	[UNK], ##스러운



Problematic for the inference and training with additional dataset.

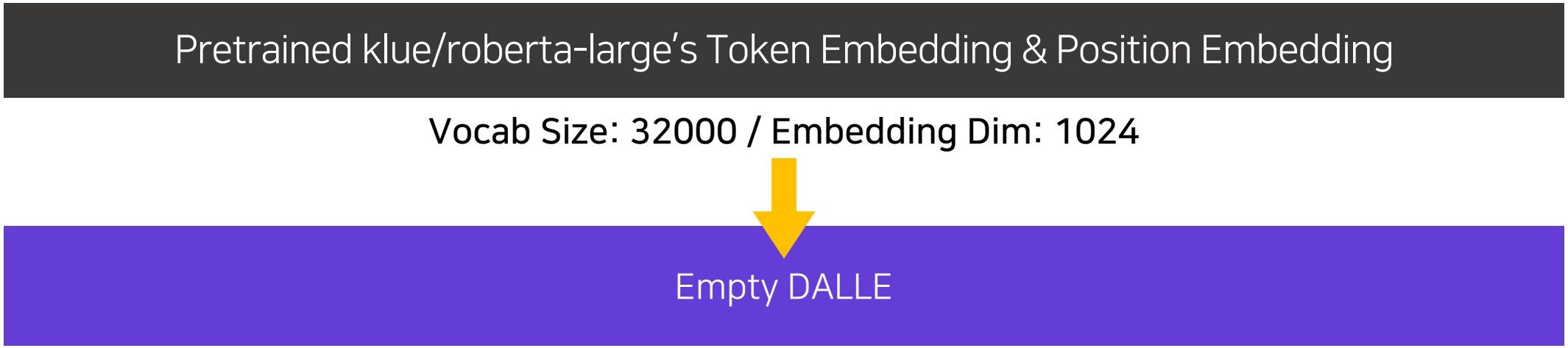
If the training dataset isn't large enough or is limited to specific domains, number of vocabularies in the trained DALLE model are insufficient.

KoDALLE



The team utilized pretrained language model's token embedding layer and position embedding layer as DALLE's text encoder.

KoDALLE



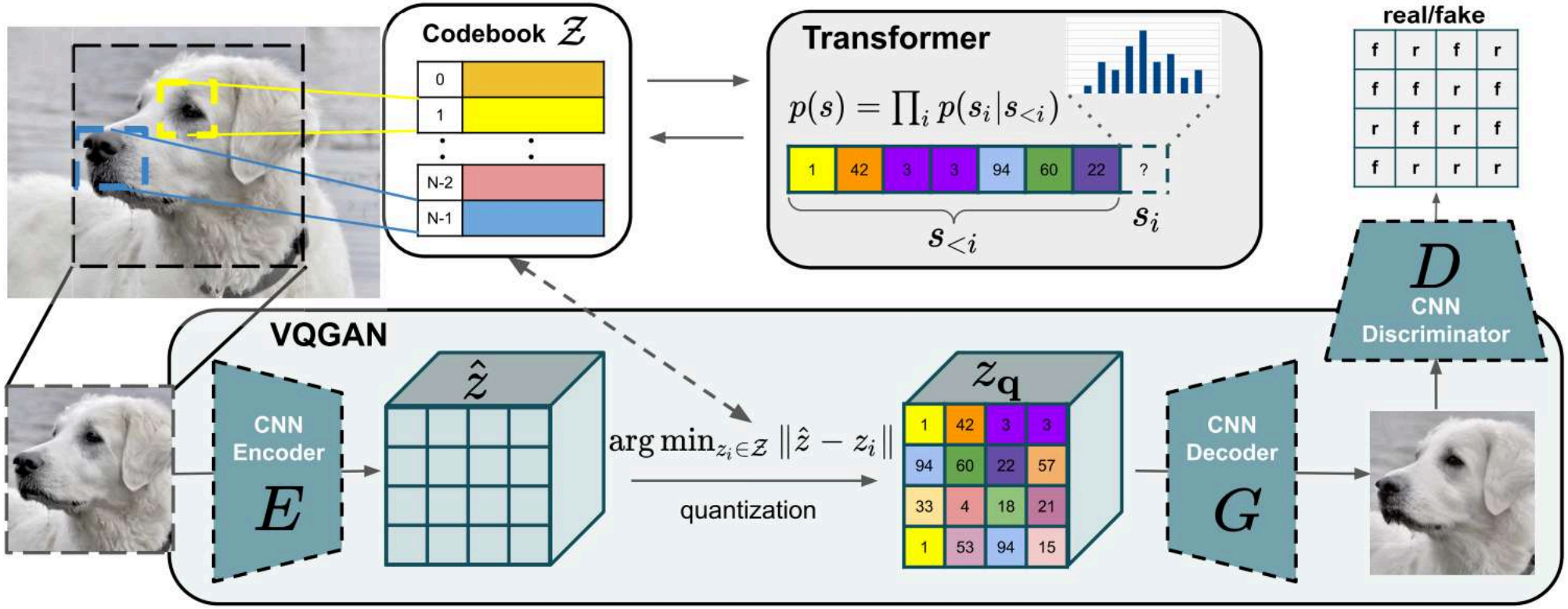
The team utilized pretrained language model's token embedding layer and position embedding layer as DALLE's text encoder.

VQGAN

VQGAN

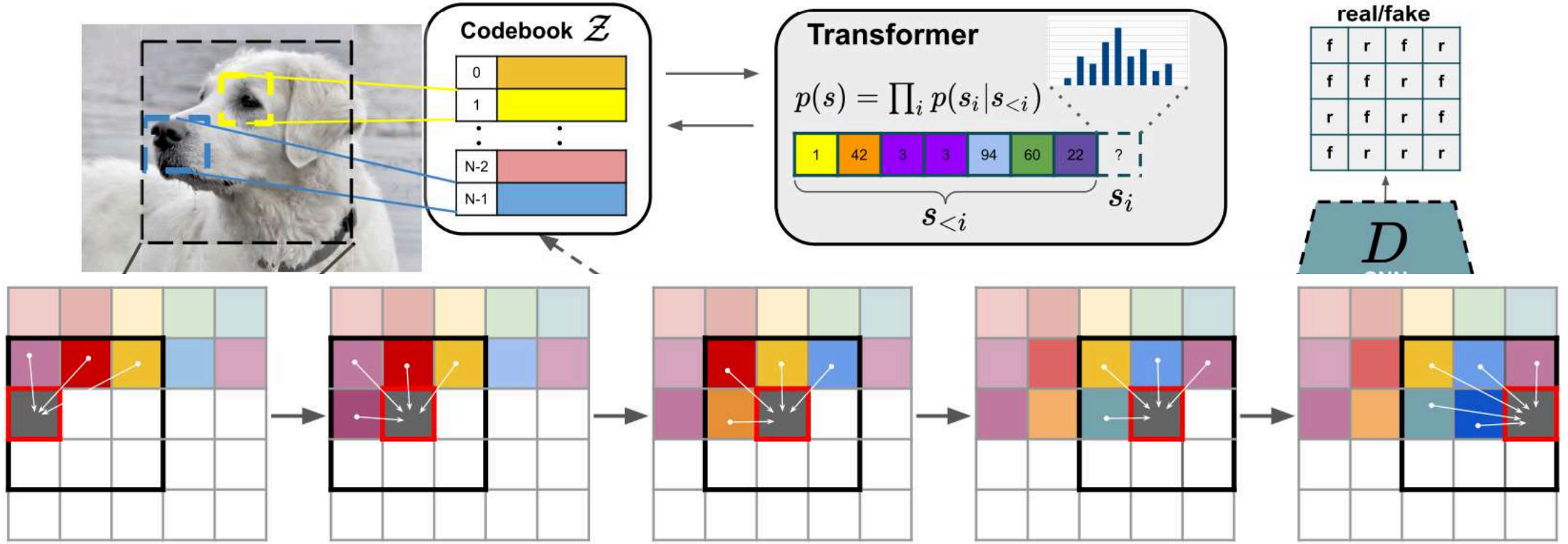
Based on Codebook of 16384 image features,
VQGAN has changed the texture, patterns and colors while maintaining the overall structure.

VQGAN



VQGAN uses both CNN and Transformers. CNN captures regional characteristics for high resolution images. Transformers captures 256 Image tokens' global interactions.

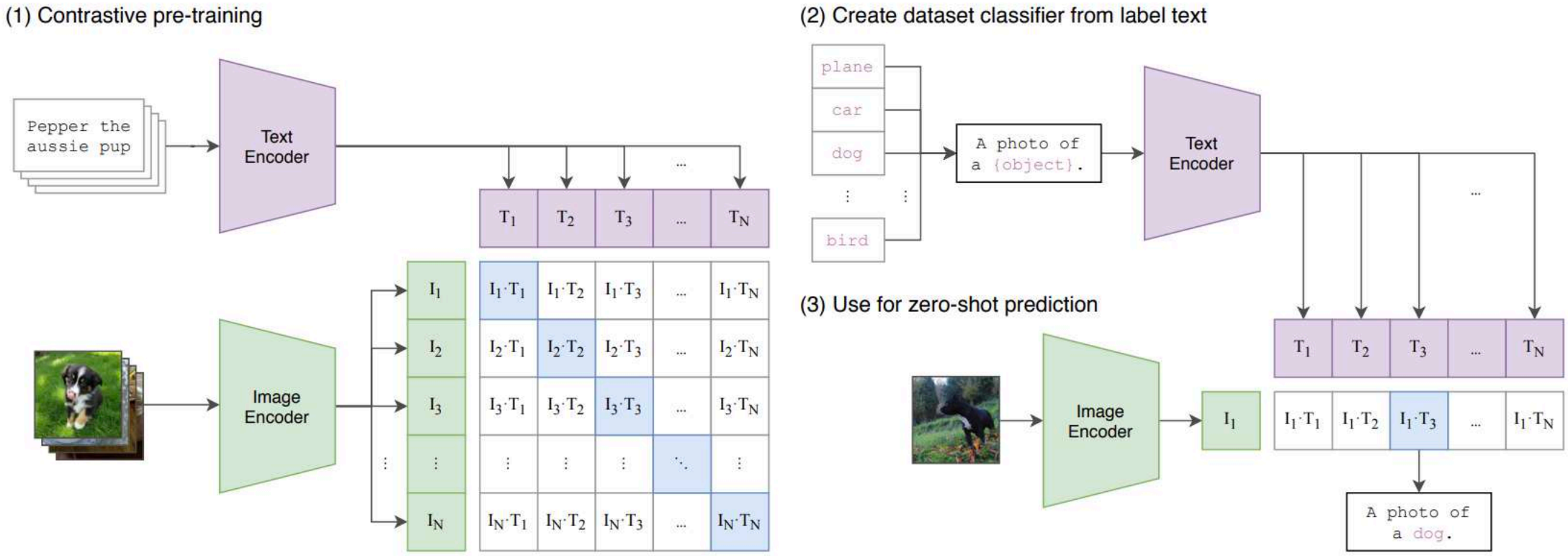
VQGAN



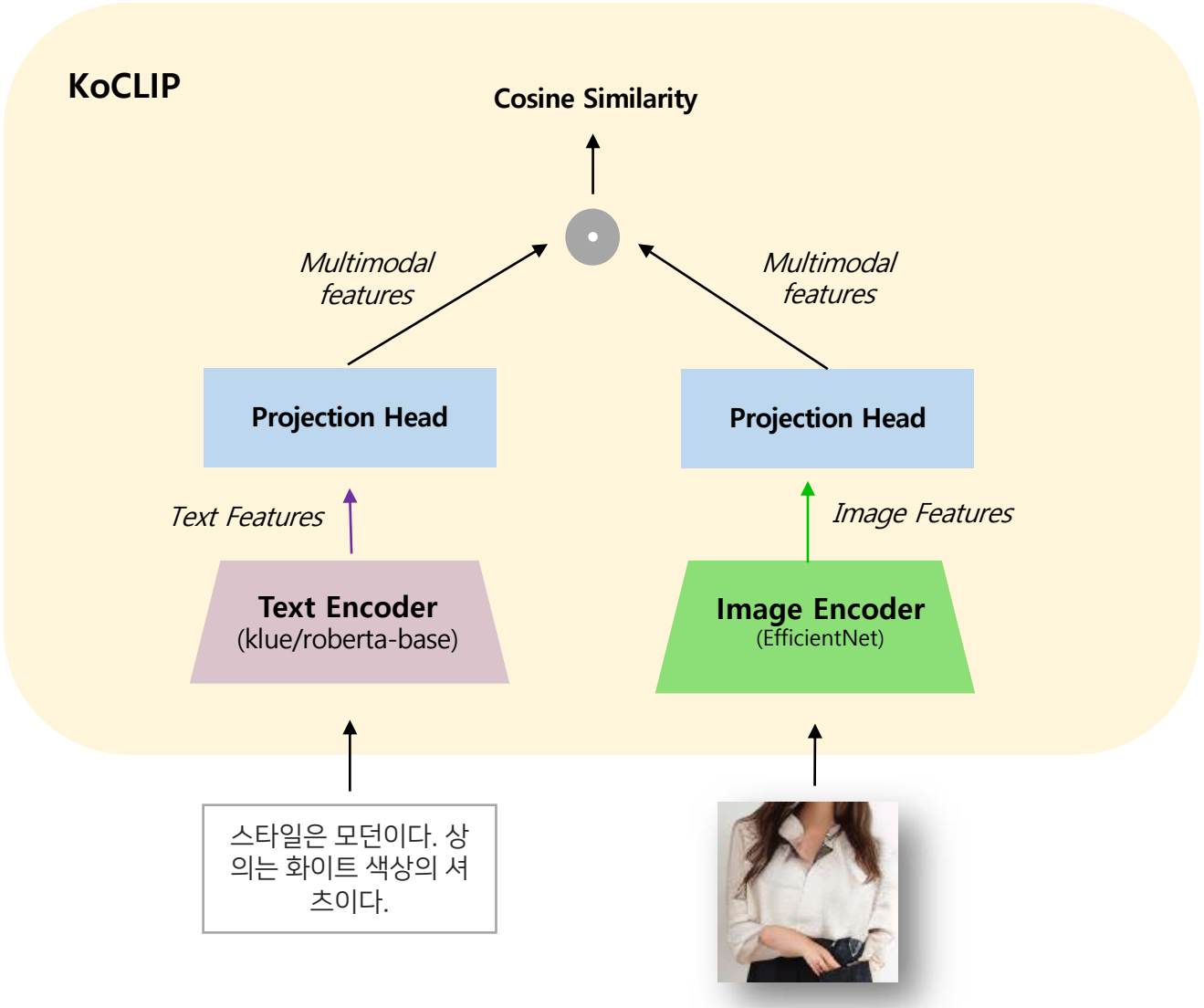
Autoregressive transformer composing image patches based on codebook.
By using sliding window technique, local features are reflected to adjacent image patches.

KoCLIP

KoCLIP



CLIP jointly trains an image encoder and a text encoder to predict the correct caption for a given image, and vice-versa.



The model was trained to maximize the cosine similarity of the correctly paired image and text embeddings while minimizing incorrect pairings.

KoCLIP & KoDALLE Significance

- Offers promising result for training from scratch on specific domains with small size dataset.
- Introduces solution for domain specific DALLE & CLIP models to be robust on input sentence.
- Recommends adequate text-to-image model size for given computation resource.
- Suggests effortless method of creating DALLE & CLIP model for own languages if pretrained l anguage model is available.

End of The Document

Appendix

Hugging Face Search models, d Models Datasets Spaces Docs Solutions Pricing

Spaces: happyface / KoDALLE like 0 Running

App Files and versions Settings

KoDALLE: Text to Fashion Image

Press CMD + Enter and Please Wait for 30 Seconds 😊

Text

상의에서 기장은 노멀이다. 색상은 화이트이다. 카테고리는 블라우스이다. 디테일에는 비대칭이다. 소매기장은 반팔이다. 소재에는 시폰이다. 프린트에는 무지이다. 넥라인은 브이넥이다. 핏은 노멀이다.

108/200

Top 4 Images Reranked with KoCLIP.

Executive Summary

DALLE Couture: Text to Fashion

Why

- Prototyping and redesigning clothes in similar manners are pain points in fashion industry.
- Yielding multiple candidates for the similar designs is time consuming.
- Loop of designing, 3D modeling, hiring fit models and analyzing is costly.

What

- **Trained and utilized VQGAN and KoDALLE for Korean text to fashion image(250x250) generation.**
- **Reranked images generated with KoDALLE by restructuring and training KoCLIP.**

How

[Phase 1: Image Generation]

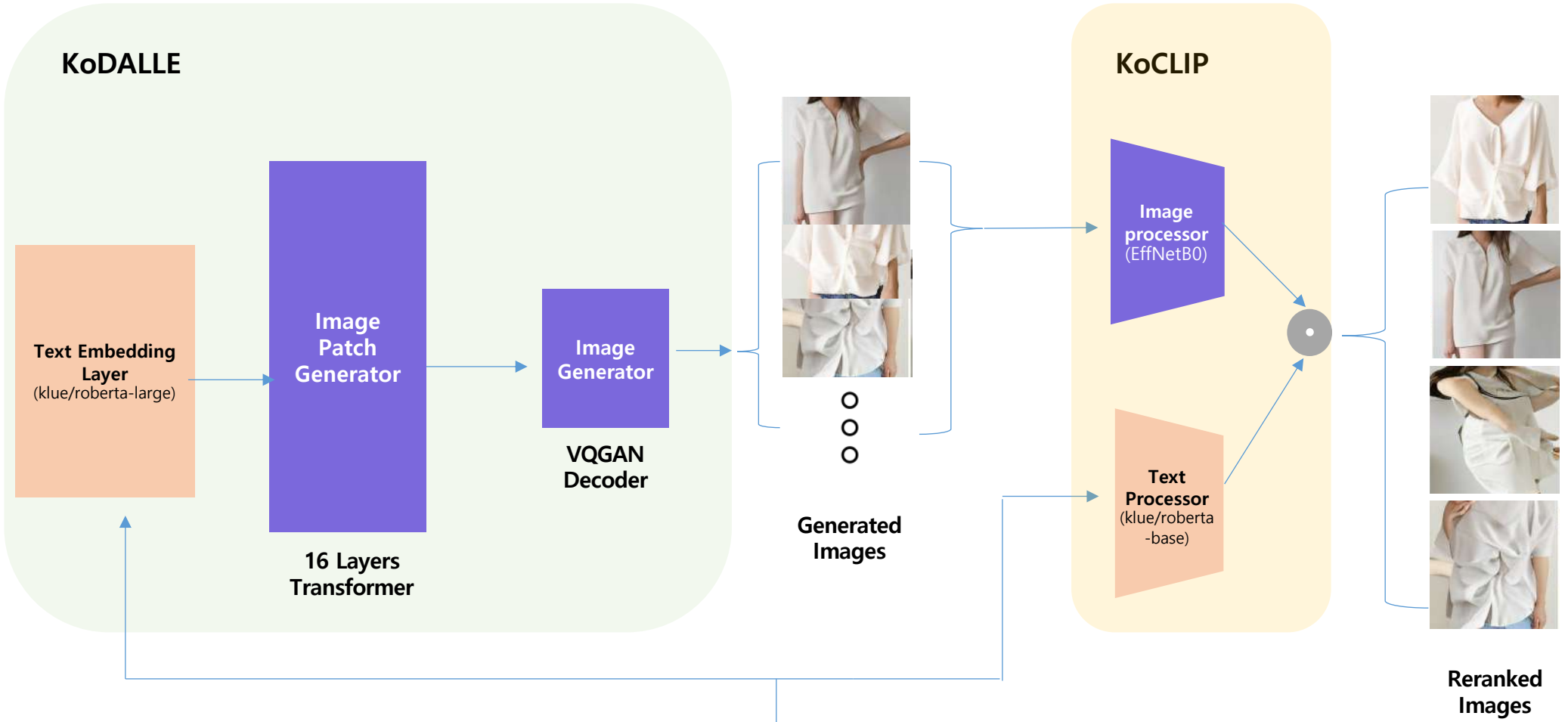
- Compared VAE, VQVAE and VQGAN for high resolution image generation.
- Trained VQGAN model from scratch on K-Fashion Image Dataset of AIHub.

[Phase 2: Text to Image]

- Compared AttentionGAN and DALLE for multimodal text to image models.
- Adapted DALLE model with Korean text embedding layer for robust inference: KoDALLE.
- Trained KoDALLE Model from scratch with K-Fashion Image Dataset with text pairs.

[Phase 3: Rerank and Deploy]

- Constructed EfficientNetB0 + klue/roberta-base for algorithmic reranking: KoCLIP.
- Deployed demo of KoDALLE model using streamlit on Huggingface Hub.


KoDALLE
Text Embedding Layer
(klue/roberta-large)
Image Patch Generator
16 Layers Transformer
Image Generator
VQGAN Decoder
Generated Images
KoCLIP
Image processor
(EffNetB0)
Text Processor
(klue/roberta-base)
Reranked Images
색상은 화이트이다. 카테고리는 블라우스이다.
넥라인은 브이넥이다. 핏은 노멀이다.
Pretrained Weights on Opensource
Trained by the team

	OpenAI's DALLE	KoDALLE of HappyFace
#Params	12 Billion	428 Million
#Layers	64 Layers	16 Layers
Computing Resource	1024 x V100 16GB	1 x V100 32GB
Text Encoder	16384 Vocab x 512 Dim BPE	32000 Vocab x 1024 Dim klue/roberta-large
Image Encoder	VQVAE	VQGAN
Optimizer	AdamW	AdamW
Learning Rate	4.5e-5	3.0e-5
Weight Decay	4.5e-3	3.0e-3
LR Scheduler	ReduceLROnPlateau	-

Q&A

Roles in Team HappyFace

김연주 @kimyeondu	• VQGAN paper review and hyperparameter optimization accordingly. • DALLE paper review and hyperparameter / input sequence optimization accordingly.
김준홍 @JoonHong-Kim	• Attached pretrained RoBERTa's text embedding layer to DALLE model. • Connected KoCLIP reranking model for KoDALLE's image generation inference process. • Experimented Contrastive Learning process for training AttnGAN.
김현수 @shawnhyeonsoo	• Developed lightweight KoCLIP model which projects text and image information into a multi-modal space to learn the similarities between texts and images.
안영진 @snoop2head	• Created training pipeline from VQGAN through KoDALLE. • Constructed inference pipeline from VQGAN through KoDALLE. • Maintained versions of dataset and preprocessed 1 million pairs of image-caption dataset.
전재영 @hihellohowareyou	• Experimented Contrastive Learning process for training DALLE. • Experimented AttnGAN for text to Image task. • Hyperparameter / input sequence label optimization.
최성욱 @jjonhwa	• Paper review for VQGAN, VQVAE for performance optimization. • Paper review for DALLE for hyperparameter optimization. • Data preprocessing for 1 million pairs of image-caption dataset.

Q&A

KoDALLE 인퍼런스 속도는?

- p100에서 1 장에 13-14초
- A100에서 16장에 12-13초

KoDALLE 파라메터 개수는?

- KoDALLE(16 Layers): 428464388(All) / 334908736(require_grad) -> VQGAN이 나머지 9M
- KoCLIP: 116202620(All) / 116202620(require_grad)

예시로써 Klue/roberta-large가 3억개 #params이고 Layer가 24개라는 것을 감안하면 충분히 연구에 사용할 수 있는 규모라고 판단함.

Q&A

DALLE를 왜 썼는지? AttnGAN과 같은 방법으로 해도 충분히 잘 나오지 않나요?

- text to image를 DALLE와 같은 transformer 계열(DALLE)과 conditional GAN(AttnGAN)의 두 가지 프로토타입을 11만장의 데이터에 대해 병렬적으로 실험을 했을 때 DALLE모델이 더 좋은 결과를 보여서 DALLE모델을 선정했습니다.
- DALLE모델이 매우 큰 모델이고 zero shot을 위해 만들어진 모델이지만 특정 specialized distribution을 학습하기 위해서는 원래 DALLE의 사이즈와 엄청나게 많은 데이터가 필요하지 않을 것이라고 판단했습니다.
- 보다 적은 레이어(16층)를 쌓아 334m개의 학습가능한 파라미터의 DALLE 모델을 60-70만장의 데이터로 여러 에폭을 통해 underfitting없이 충분히 모델을 학습 시킬 수 있을 것이라고 생각했습니다.
- DALLE 라는 모델이 zero shot으로 CUB dataset과 같은 specialized distribution을 가지는 구체적인 영역에서는 좋은 성능을 보이지 못했는데 특정 영역에 대한 학습을 통해 이런 specialized distribution에서도 좋은 성능을 보일 수 있음을 보였다는 점에서 의의가 있다고 생각합니다.

Q&A

VQGAN vs VQVAE

- Rich Codebook과 Transformer, Discriminator를 추가하였다.
- Rich Codebook과 Transformer를 바탕으로 High-Resolution을 표현할 수 있게 만들었다
- Discriminator를 바탕으로 perceptual loss를 추가함으로서 compressed rate가 증가함에 따른 perceptual quality를 유지할 수 있다.

KoDALLE를 진행하면서 어려웠던 점은?

- Domain 지식이 없었다.
- 우리의 과정은 Task Specific하기 때문에 짧은 시간 안에 다양한 Paper와 연구결과 비교 및 성능 비교하기가 어려웠다. AttnGAN이 더 좋을수도 있고, DALLE가 더 좋을 수도 있고.
- “왜?”라는 질문의 연속: vqgan이 dvae보다 좋다. 왜 좋을까? bpe보다 pre-trained weight tokenizer를 쓰는게 왜 좋을까? 등등