GAN을 이용한
텍스트-이미지, 텍스트-음악
프로그램 구현

전주 ICT 이노베이션 스퀘어 온라인코딩교육 1조
양명훈 김아영 송경민 오한나 허지원

목차

1. 프로젝트 제안 배경 및 목표

- 글을 입력하여 어울리는 그림, 음악이 나오면 좋을 것
- Text to Image(텍스트의 이미지화)
- Text to Music(텍스트의 음악화)

2. 프로젝트 프로세스

- 주제 선정 및 데이터 수집
- 탐구 / 탐색 팀으로 나뉘어 스터디 및 정보 검색
- 탐구 / 탐색 정보 공유
- 작곡/ 이미지 팀으로 나뉘어 코드 분석 및 이해
- 최종 산출물(리포트) 작성

3. 프로젝트 요약 - 음악

- 음악 생성 모델(Generate Music with AI)은 언어 생성 모델과 유사
  비지도 학습으로 데이터 학습, 아웃풋 생성
- 텍스트를 음악으로 만드는 프로그램을 구현하기 위해선
  input 과정이 하나 더 추가되어야 함
- 이번 프로젝트에서는 목표 달성을 위한 사전 배경 지식 학습에 중점

향후 발전 방향

1. input 텍스트를 5가지 분위기(태그)로 분류
2. 분위기가 맞는 음악을 학습시킨 모델 5가지를 만듦
3. 모델을 API로 만들어 배포
   - 장르별로 태그가 넘어오면 음악을 만들어 주는 식

- 현재는 영어만 가능, 한글 입력시 결과물 나오는 프로그램 구축
  (작곡, 이미지 공통사항)

3. 프로젝트 요약 - 이미지

- 이미지 생성 모델을 이해하기 위한 사전 지식 학습
  GAN, DALL-E, VQ-VAE, U-Net
- 사전 훈련된 모델을 이용하여 이미지 생성해보기

향후 발전 방향

- pyTorch를 텐서플로우로 변환 혹은 pyTorch로 코딩
- text-to-image API 배포
- GCP(Google Cloud Platform)자료 활용 업그레이드
- 현재는 영어만 가능, 한글 입력시 결과물 나오는 프로그램 구축
  (작곡, 이미지 공통사항)

4. 단계별 내용 - 작곡

- 음악 생성 모델(Generate Music with AI)은 언어 생성 모델과 유사성을 띔
- 비지도 학습으로 데이터 학습, 아웃풋 생성

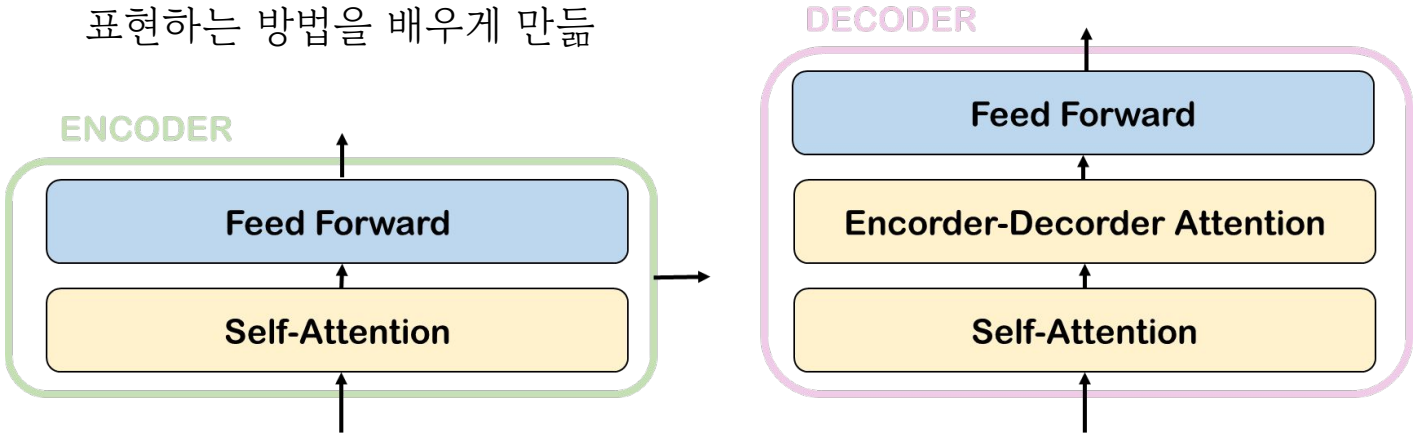
필요 배경 지식

- AutoEncoder
- GAN(Generative Adversarial Network)
- Transformer
- Attention
- 다음 페이지에서 자세한 설명

4. 단계별 내용 - 작곡

AutoEncoder : encoder(차원축소) + decoder (복원)

- 비지도학습 모델, 입력과 출력 크기 동일
- 단순히 입력을 출력으로 복사하는 방법 학습
- 다양한 방법으로 네트워크에 제약을 가해 작업을 어렵게 만듦
- 단순히 입력을 출력으로 바로 복사하지 못하도록 막고 data를 효율적으로 표현하는 방법을 배우게 만듦



4. 단계별 내용 - 작곡

GAN (Generative Adversarial Network) : 적대적 신경생성망

- 생성자, 구별자의 대립 시스템, 경쟁을 통해 서로의 성능을 점진적 개선
- Generator(생성자) + Discriminator(구별자)

Transformer : 시퀀스-투-시퀀스 과제 수행 위한 모델

- RNN 사용하지 않음
- Encorder-Decoder 구조
- 기존 seq2seq 구조인 Encorder-Decoder를 따르지만, Attention으로 구현한 모델

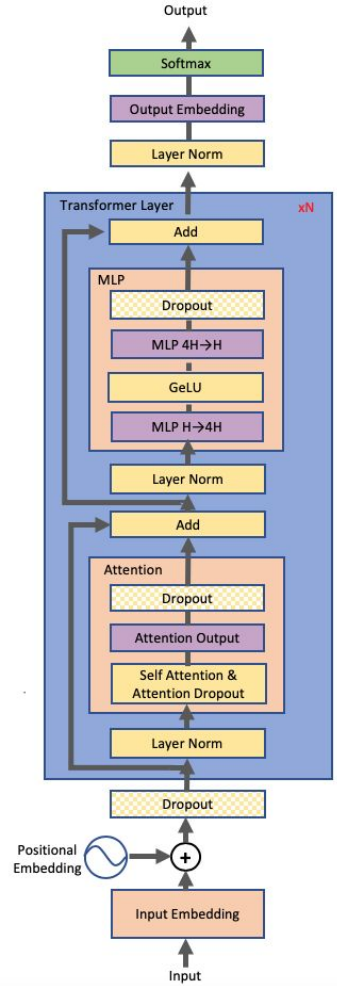
Attention

- RNN 단점 보완(고정 크기의 벡터 -> 정보 손실, 기울기 손실)
- 출력 단계별로 컨텍스트 백터 생성 방법 학습
- 모델이 입력 시퀀스와 생성 결과를 통해 무엇에 집중(Attention) 할 것인지 학습

4. 단계별 내용 - 작곡

GPT (Generative Pre-trained Transformer)

- Attention을 사용한 언어 모델
- 주어진 단어로부터 그 다음에 나올 단어(Token)를 예측
- 다음 단어를 예측하려면 글 전체 맥락을 봐야 하는데,
  음악에서도 마찬가지로 곡 전체 분위기를 보고 다음 Note 예측
- Transformer의 Decoder 구조와 유사



4. 단계별 내용 - 작곡

MuseNet

- OpenAI 개발
- 10가지 악기를 이용, 4분 내외의 음악을 생성하는 AI모델
- Input으로 6개의 Note를 입력하면 선택한 장르의 음악을 생성
- GPT-2 사용
- Sparse Transformer(긴 구조를 하나로 기억) 사용

4. 단계별 내용 - 작곡

Optimus-VIRTUOSO

- Github 오픈 코드
- 다중 악기를 이용
- MuseNet과 유사한 MIDI 이벤트 표현
- GPT-3 사용
- 일반 Music AI이기 때문에 독창적인 음악은 작곡할 수 없음

4. 단계별 내용 - 작곡

사용 모델 및 모듈

1. minGPT
   - GPT를 이해하기 쉽게 만든 학습용 모델
   - 파라미터만 바꾸면 GPT-1, GPT-2, GPT-3 모두 표현 가능
   - Andrej Karpathy(**前** 테슬라 AI 개발 총괄)가 만듦
   - PyTorch 사용

2. TMIDIX
   - 미디 파일 관련 모듈

결과물

4. 단계별 내용 - 이미지

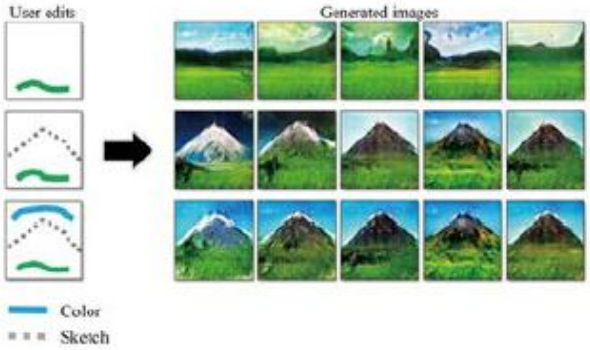
- **Text to Image(텍스트의 이미지화)**
- 필요사전지식: GAN, DALL-E, VQ-VAE, U-Net
  → 사전학습된 모델을 이용한 text-to-image 결과물

Generative Adversarial Network (GAN) 적대적 생성신경망

- 서로 대립하는 두 시스템의 경쟁을 통해 학습하는 방법론
- 주요 응용분야: 이미지 생성 복원, 동작 흉내 인공지능, 신약 개발 등
- GAN의 종류: StyleGAN, SGGan, DCGAN 등

4. 단계별 내용 - 이미지

[그림 1-2] GAN을 활용한 이미지 생성 iGAN



※ 자료 : iGAN https://github.com/junyanz/iGAN

[그림 1-3] GAN을 활용한 이미지 복원 : 화질이 낮은 이미지 (좌) GAN을 활용한 복원 (우)

※ 자료 : Ian Goodfellow, NIPS 2016 Tutorial: Generative Adversarial Networks

4. 단계별 내용 - 이미지

이러한 경쟁적 학습 지속되면.....

4. 단계별 내용 - 이미지

어느 지폐가 진짜? 위조?

확률은 50:50

4. 단계별 내용 - 이미지

DALL-E

- OpenAI 개발
- 자연어처리와 컴퓨터 비전을 결합하여 텍스트에서 이미지를 생성할 수 있는 AI모델
- 텍스트-이미지 쌍의 데이터세트를 사용하여 텍스트 설명에서 이미지를 생성하도록 훈련된, 120억 개의 매개 변수로 이루어진 Transformer 기반의 GPT-3의 확장 형태

- 초현실주의의 화가 살바도르 달리(Salvador Dali)와 로봇 애니메이션 속 로봇 캐릭터 월-E(WALL-E)에서 영감을 받아 *DALL-E* 라는 이름이 됨
- 동물과 사물의 의인화된 버전 생성, 관련 없는 개념을 그럴듯하게 결합, 텍스트 렌더링, 기존 이미지 변형 등 다양한 기능

4. 단계별 내용 - 이미지

DALL-E

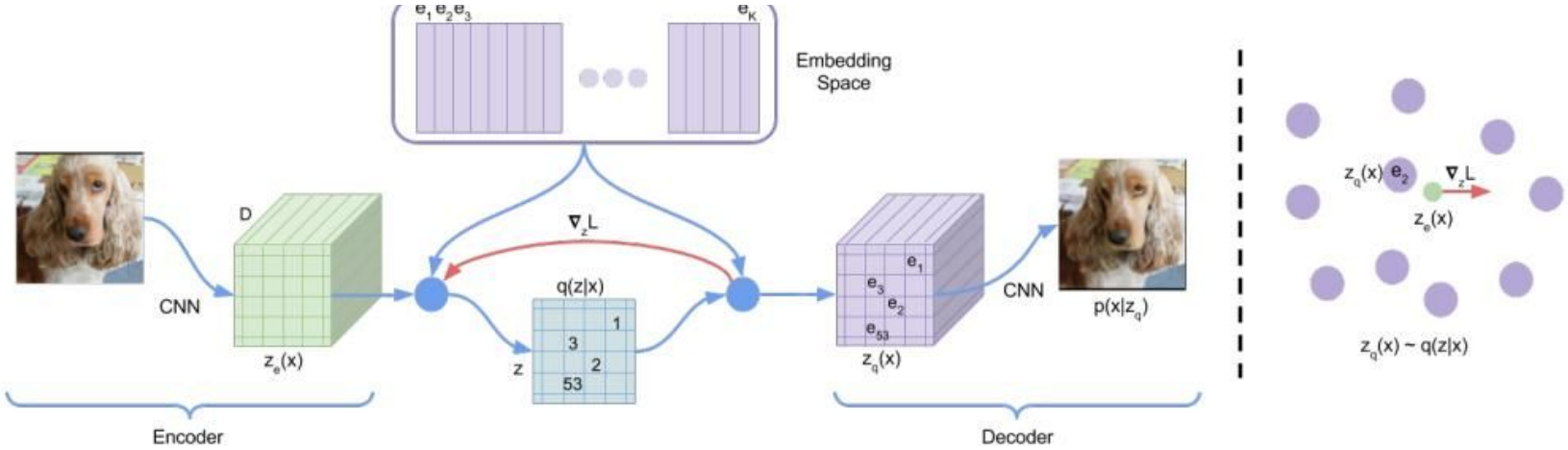
AI 생성 이미지

아보카도 모양의 안락의자 **(text-to-image)**

동일한 고양이 사진과 여러가지 스케치 **(image-to-image)**

4. 단계별 내용 - 이미지

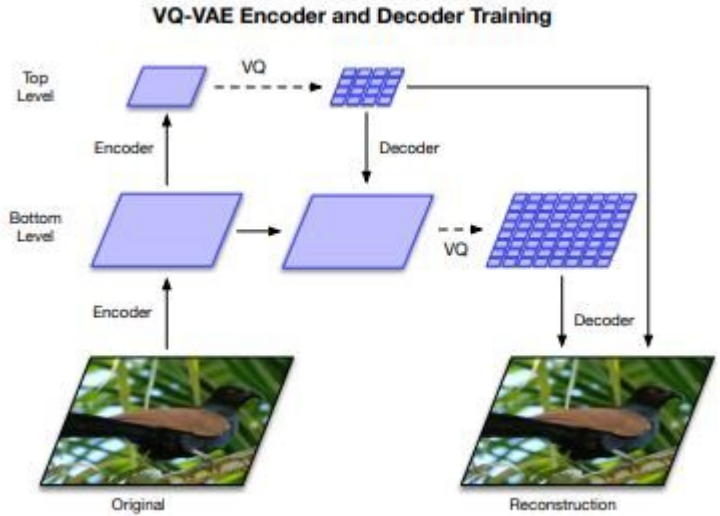
VQ-VAE Vetor Quantised-Variational AutoEncoder



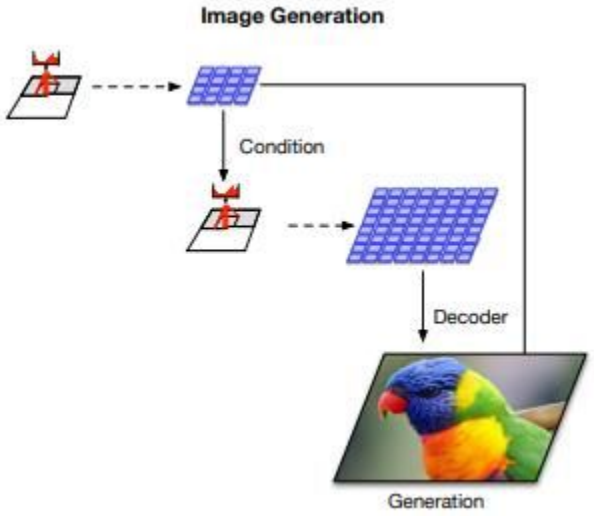
VQ-VAE의 구조

4. 단계별 내용 - 이미지

VQ-VAE2 Generation Diverse High-Fidelity images with VQ-VAE2



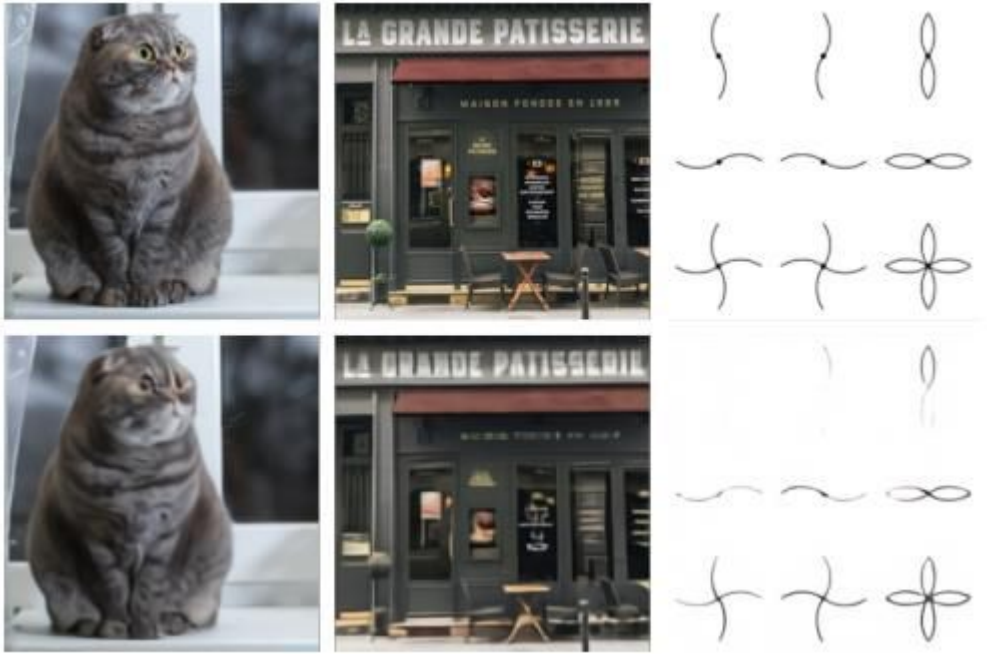
Stage1
VQ-VAE training



Stage2
image Generation

4. 단계별 내용 - 이미지

DALL-E Methodology



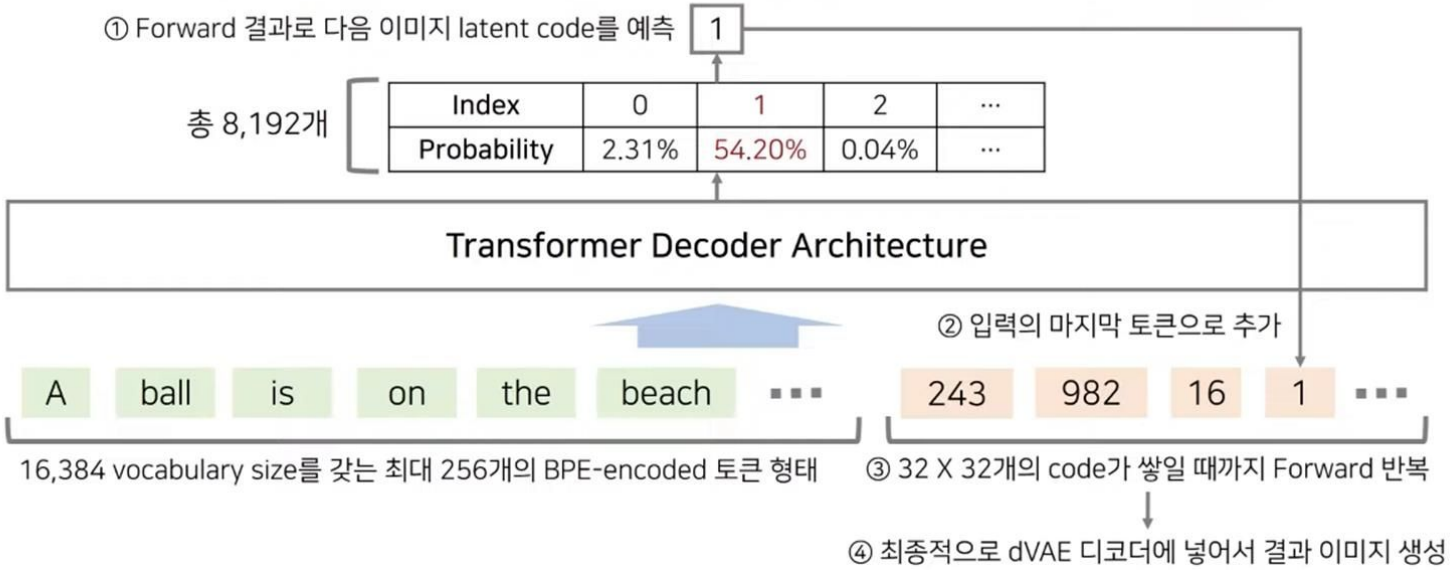
원본

VQ-VAE 결과

4. 단계별 내용 - 이미지

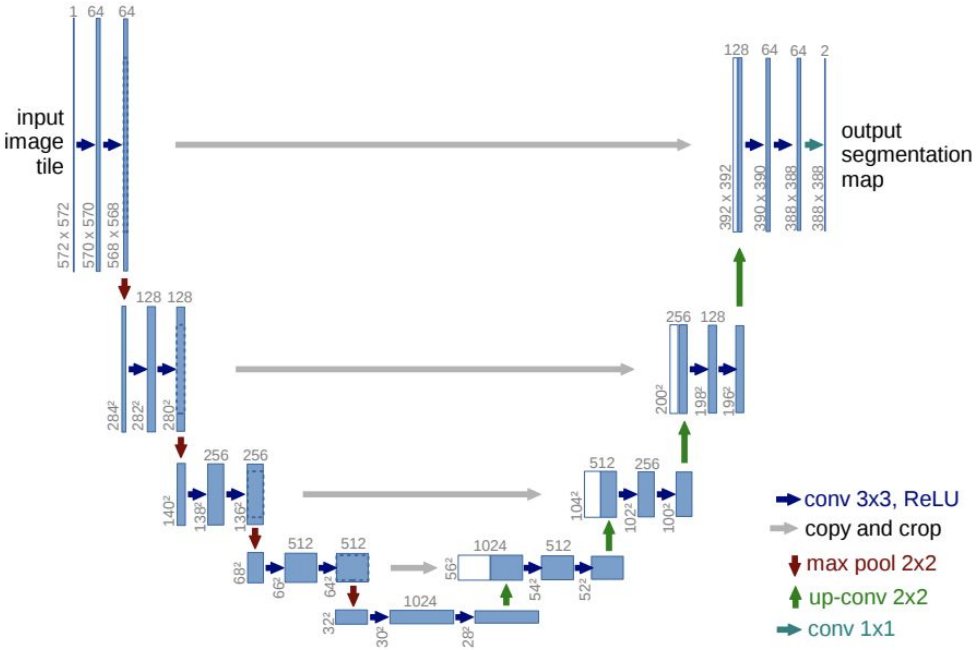
DALL-E Methodology

- 먼저 text token들이 최대 256개 들어가고, 이어서 image token들이 최대 1,024개 입력될 수 있습니다.
- 모델을 사용할 때는 text만 넣거나 text + image(rectangular region)를 넣어서 결과 이미지를 생성할 수 있습니다.



4. 단계별 내용 - 이미지

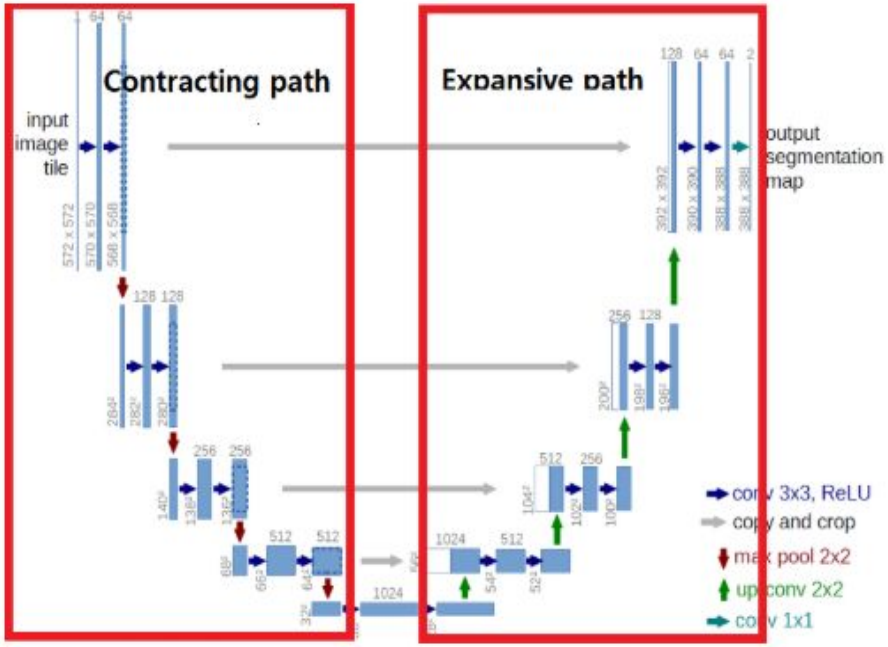
U-Net Convolutional Networks for Biomedical Image Segmentation

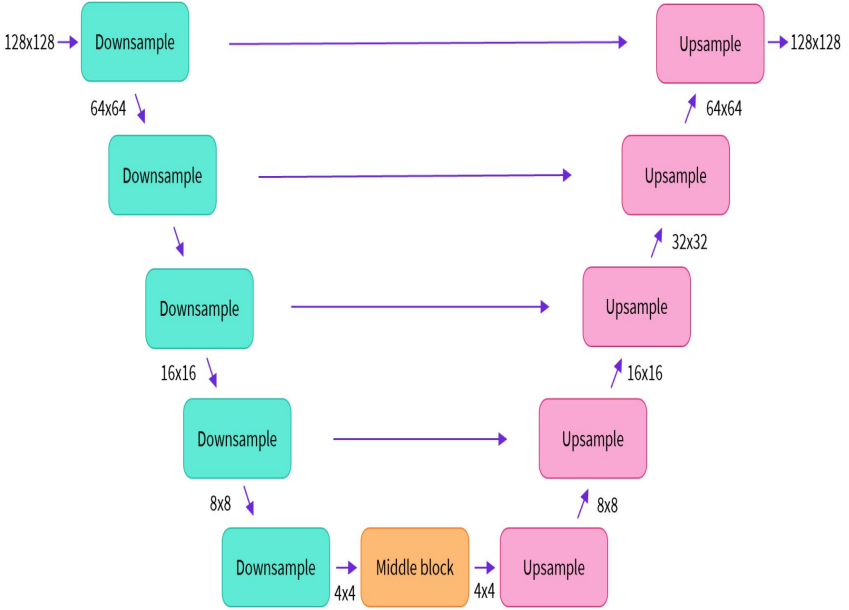


- U 모양처럼 생겨서 U-Net
- Biomedical 분야에서 적은 데이터를 가지고도 더욱 정확한 이미지 분할(Image Segmentation)을 내기 위해 Fully-Convolutional Network(FCNs) 기반으로 확장한 End-to-End 방식의 모델

4. 단계별 내용 - 이미지

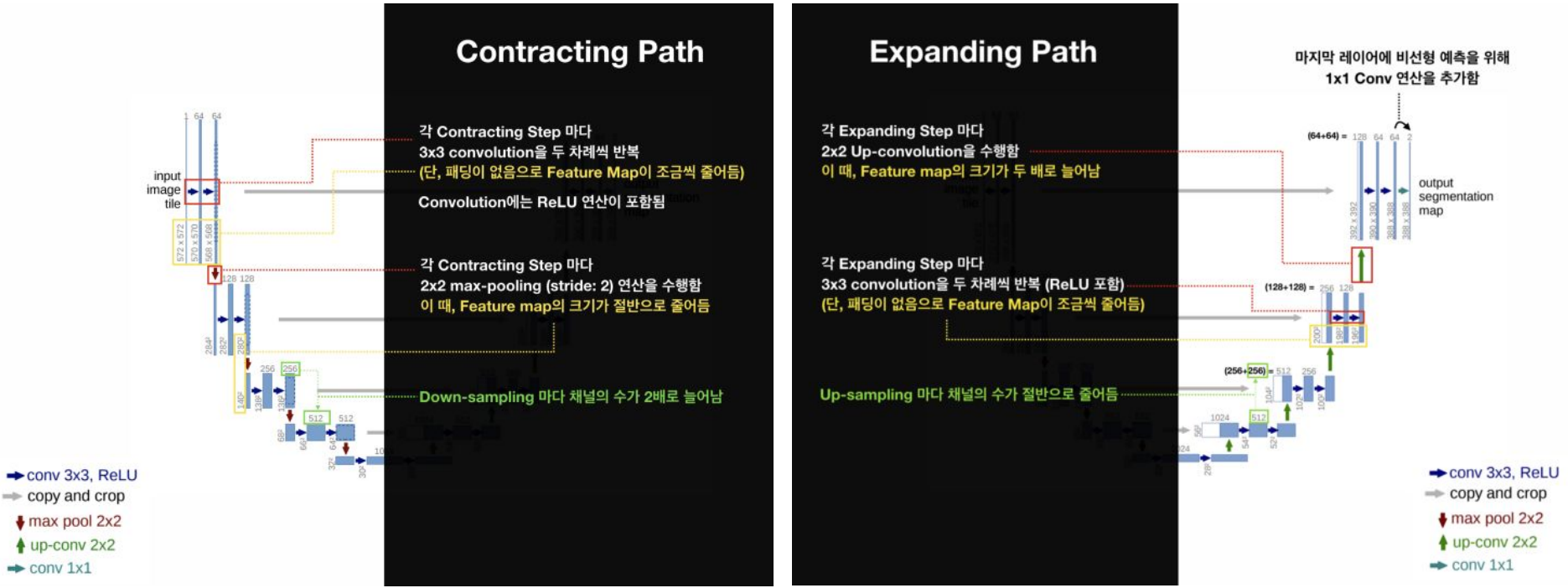
U-Net 구조: 수축 경로와 확장 경로





4. 단계별 내용 - 이미지

U-Net 구조: 수축 경로와 확장 경로



4. 단계별 내용 - 이미지

Stable Diffusion 모델

- LAION-5B 데이터베이스의 하위 집합에서 512x512개의 이미지 학습(LAION-5)
- U-Net과 동결된 CLIP ViT-L/14(사전 훈련된) 텍스트 인코더를 사용하는 잠재 확산 모델이다.
- 각 단계에서 이미지를 약간 노이즈 제거 하는 방법을 예측하도록 훈련된다. 일정 수의 단계 후에 표본을 추출한다.

→ 사전학습된 Stable Diffusion모델로
text-to-image Go

4. 단계별 내용 - 이미지

Stable Diffusion 모델

```
prompt = "a photograph of an astronaut riding a horse"
```

우주 비행사가 말을 타고 있는 사진, 프롬프트 입력

5. 프로젝트 후기

- **프로젝트 진행 과정 중 애로사항**

1. 대부분의 코드가 pyTorch로 이루어져 있어서 코드 해석 난항
   - 이미지, 작곡 공통적 문제
2. 뮤즈넷 가이드 코드 (의존성 문제)
   - tensor2tensor 라이브러리 설치 오류, tensorflow v1에 의존함
   - Colab에서는 tensorflow v1 실행 불가능(지원하지 않음)
3. 뮤즈넷 API 이용 코드
   - 여러가지 장르를 선택할 수 있다는 장점이 있으나,
   - API를 사용하다보니 Request, Response 방식으로 미디 파일을 받기 때문에 내부 구조를 알 수가 없어서 코드적 이해가 어려움
4. 결과물
   - 음악 완성도가 떨어짐, 듣기 좋지 않음

5. 프로젝트 후기

- **프로젝트 진행 과정 중 문제점**
  - 인종, 성별 등 차별과 편견을 학습하는 AI

- **느낀 점**
  - 공부하면 할 수록 배울 개념이 많아져 더욱 어려워짐
  - pyTorch로 작성된 코드가 많아 pyTorch 공부의 필요성
  - 논문 읽는 연습 필요
  - 당일 진행한 내용은 당일 정리 및 기록하는 습관 들이기 (잊어버리는 해프닝 발생)

6. 참고 문헌

작곡
https://arxiv.org/pdf/1706.03762.pdf
MuseNet
Generating Piano Music with Transformer
1) 어텐션 메커니즘 (Attention Mechanism) - 딥 러닝을 이용한 자연어 처리 입문
1) 트랜스포머(Transformer) - 딥 러닝을 이용한 자연어 처리 입문
Generative Modeling with Sparse Transformers
Attention Is All You Need(transformer) paper 정리 | by 정민수
밑바닥부터 이해하는 어텐션 메커니즘(Attention Mechanism)
GitHub - karpathy/minGPT: A minimal PyTorch re-implementation of the OpenAI GPT (Generative Pretrained Transformer) training
https://github.com/asigalov61/Tegridy-MIDI-Dataset/blob/master/Advanced_MIDI_Channel_Splitter.ipynb
GPT (Generative Pre-trained Transformer) 학습시키기
https://github.com/asigalov61/tegridy-tools
Attention Mechanism 시각화
[GAN] 오토인코더와 GAN을 사용한 표현 학습과 생성적 학습

6. 참고 문헌

이미지

적대적 생성신경망(Generative Adversarial Network)의 소개와 활용 현황
[외부기고] [새로운 인공지능 기술 GAN] ② GAN의 개념과 이해
DALL·E: Creating Images from Text
DALL-E - Wikipedia
[AI 모델 탐험기] #18 그림 그리는 AI, DALL-E | by AI Network
The Annotated Diffusion Model
GitHub - huggingface/diffusers: 🤗 Diffusers: State-of-the-art diffusion models for image and audio generation in PyTorch
CompVis/stable-diffusion-v1-4 · Hugging Face
Diffusion Models | Paper Explanation | Math Explained
AI는 어쩌다 편견과 혐오를 배웠을까